unibz
Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 85/ 2021

Exponential Tilting for Zero-inflated Interval Regression with Applications to Cyber Security Survey Data

Cristian Roner, Claudia Di Caterina,
Davide Ferrari

Exponential Tilting for Zero-inflated Interval Regression with Applications to Cyber Security Survey Data

Cristian Roner[†], Claudia Di Caterina[†] and Davide Ferrari[†]

[†]*Free University of Bozen-Bolzano, Piazza Università 1, 39100, Bozen-Bolzano, Italy.*
(*e-mail: cristian.roner@unibz.it,claudia.dicaterina@unibz.it,davide.ferrari2@unibz.it*)

Abstract
Non-negative ordered survey data often exhibit an unusually high frequency of zeros in the first interval. Zero-inflated ordered probit models handle the excess of zeros by combining a split probit model and an ordered probit model. In the presence of data violating distributional assumptions, standard inference based on the maximum likelihood method gives biased estimates with large standard errors. In this paper, we consider robust inference for the zero-inflated ordered probit model based on the exponential tilting methodology. Exponential tilting selects unequal weights for the observations in such a way as to maximise the likelihood function subject to moving a given distance from equally weighted scores. As a result, observations that are incompatible with the assumed zero-inflated distribution receive a relatively small weight. Our methodology is motivated by the analysis of survey data on cyber security breaches to study the relationship between investments in cyber defences and costs from cyber breaches. Robust estimates obtained via tilting clearly show an effect of the investments in reducing the amount of the loss from a cyber breach.

Keywords
Zero-inflation, Exponential tilting, Interval regression, Cyber security, Survey data.

JEL classification numbers
C1, C13, C83, D24, D25

I. Introduction

Cyber attacks are malignant assaults launched against single computers or a computer network in order to gain access or make use of an asset such as customer data, patents, product specifics etc. In today's economy cyber attacks are becoming increasingly common threats and successful breaches are detrimental to firms, households and entire economic sectors. Although the study of the economic impact of cyber security breaches has gained relevance in recent years (e.g., see Romanosky 2016, Biancotti 2017, Eling and Wirfs 2019), there is still a substantial lack of empirical research helping economists identify the factors associated with the costs of cyber security attacks. In this regard, most of the existing empirical work relies on survey data collected at the firm level. For example, in this paper we consider data from the United Kingdom Cyber Security Breaches Survey[1] (CSBS). The CSBS is one of the first systematic data collection initia-

[1] https://www.gov.uk/government/collections/cyber-security-breaches-survey

tives to directly assess the exposure of firms to cyber attacks and raise awareness at the country level.

Response variables of interest in cyber security surveys often have ordinal nature and take interval values. In the CSBS data, for instance, the cost associated with a cyber attack (in 1,000 GBP) takes interval values $[0, 0.5)$, $[0.5, 1)$, $[1, 5)$, etc. The distribution for the cost variable reported in Table 4 shows particularly high frequency in the first interval class containing the value zero. A popular econometric approach for ordinal survey responses is the ordinal probit (OP) model. A useful special case of the OP model for interval censoring is obtained by fixing the threshold parameters determining response categories; this is often referred to as interval regression (IR) model. However, OP and IR models are not useful when the sample distribution of the response exhibits a high frequency for the category counting zero. Ignoring the zero-inflation feature of the data is known to produce inaccurate estimates; e.g., see Harris and Zhao (2007); Brown et al. (2015).

There is a vast literature on zero-inflated models in statistics and econometrics; e.g., see the early approaches described in Mullahy (1986), Pohlmeier and Ulrich (1995), Gurmu and Trivedi (1996), Hall (2000) and Min and Agresti (2005). Appealing models in this class are the two-stage type models since they relax the assumption that the zeroes and strictly positive observations come from the same data-generating process. This enables one to assess separately, within a regression context, the impact of independent variables on zero and non-zero outcomes. Harris and Zhao (2007) develop the zero-inflated ordered probit (ZIOP) model consisting of a combination of a split probit model to describe the inflation at zero and an ordered probit model. In recent years, ZIOP models have proved useful in a variety of applications; e.g., see Downward et al. (2011), Jiang et al. (2013), Bagozzi et al. (2015) and Tan and Yen (2017). Gurmu and Dagne (2012) consider Bayesian estimation with applications to tobacco use data, while Das and Das (2018) develop a zero-inflated semi-parametric ordinal model with a non-linear link between an ordinal response variable and a set of covariates. Brown et al. (2015) consider the zero-inflated interval regression (ZIIR) model, a special case of the ZIOP model with boundary parameters for each ordinal category fixed, and apply it to primary care survey data.

Although ZIOP and ZIIR models are more realistic compared to their traditional counterparts (OP and IR, respectively), inference is inaccurate when the model is mis-specified, i.e. when the data distribution deviates to some extent from the assumed nominal model. Estimation is currently mostly limited to the maximum likelihood (ML) methodology, which has poor performance in the presence of model mis-specifications. Motivated by these limitations, we develop a robust estimator for zero-inflated models using the general framework of exponential tilting. Data tilting involves replacing uniform data weights with more flexible weights to render parametric procedures more robust to model mis-specifications and obtain greater fitness. In addition, the tilted weights produce a natural ordering of the observations with respect to their compatibility with the assumed model, thus enabling outlier detection. Here, we propose to obtain weights in such a way as to maximise the likelihood function subject to moving a given distance (Kullback-Leibler divergence) from equally weighted scores. The use of tilting has been previously investigated by Hall and Presnell (1999), Choi et al. (2000), Critchley and Marriott (2004), Camponovo and Otsu (2012), Genton and Hall (2015), and Ferrari and Zheng (2016), among others. Although our development focuses on ZIIR models due to the nature of our application, an extension of the methodology for general ZIOP models is straightforward using analogous principles.

The exponential tilting methodology for the ZIIR model is applied to the 2018-2019 CSBS data. To the best of our knowledge, this is the first work applying a zero-inflated regression approach to cyber security data. The ZIIR model is particularly suited for this application since it allows us to simultaneously estimate the probability of sustaining a loss and a regression model for the sustained cost when this actually occurs. Our analysis shows that the normality assumption underlying the latent processes in the first- and second-stage probit models are violated for the CSBS data. This makes our tilting estimation technique useful to better understand the determinants underpinning costs of cyber breaches. Such insights are relevant for managing firms and institutions, formulating policies and for formulating economic theories. In particular, the relation between the costs of cyber breaches and the investments in cyber defence deserves to be further investigated.

The paper is organized as follows. In Section II, we introduce the ZIIR model and describe the proposed exponential tilting estimation method. In Section III, we present Monte Carlo experiments assessing the performance of the robust method in comparison to the standard ML approach. In Section IV, we apply our methodology to the CSBS data and discuss the results. In Section V, we conclude and provide indications for future research.

II. Methodology

Zero-inflated interval regression

The CSBS survey report monetary loss from cyber attacks on an interval scale; thus the real loss may be naturally viewed as a latent variable. Let Y_i^* be the actual (unobservable) monetary loss sustained by company i. Although we cannot directly observe Y_i^*, we can observe the ordered categories

$$Y_i = \begin{cases} 0, & \text{if} \quad Y_i^* < \gamma_1, \\ 1, & \text{if} \quad \gamma_1 < Y_i^* < \gamma_2, \\ \vdots & \\ K, & \text{if} \quad Y_i^* > \gamma_K, \end{cases}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$ is a vector of given threshold values. For the unobservable response of interest, Y^*, we assume a two-stage selection model. Note that we focus on the case where the thresholds are fixed; this is a special case of the ZIOP model called zero-inflated interval regression (ZIIR). However, the inference methods developed in the reminder of the paper are also appropriate in the case of unknown thresholds, provided that adequate parameter restrictions ensuring model identifiability are introduced.

The unobservable monetary loss for an individual organization is modelled in two stages. In the first stage, we specify the probability that an actual loss occurs using the following multivariate latent model. Let S_i be a binary variable indicating whether a loss occurs for the ith individual; particularly, $Y_i^* = 0$ if $S_i = 0$ and $Y^* > 0$ if $S_i = 1$. The dependence between S_i and a $p \times 1$ vector of covariates $\mathbf{x}_i$ observed for firm i is given by the probit model:

$$\begin{aligned} S_i^* &= \mathbf{x}_i^\top \boldsymbol{\beta}^{(1)} + U_i^{(1)}, \\ S_i &= I(S_i^* > 0), \quad i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta}^{(1)}$ is a $p \times 1$ vector of unknown parameters, $(U_1^{(1)}, \ldots, U_n^{(1)})$ are random errors assumed to follow a n-variate normal distribution with mean $E(\boldsymbol{U}^{(1)}) = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}^{(1)} = \sigma^{(1)}\boldsymbol{I}_n$.

The second stage of our model specifies how much a firm loses when such a loss does occur. Particularly, we assume the latent regression model:

$$Y_i^* = \mathbf{z}_i^\top \boldsymbol{\beta}^{(2)} + U_i^{(2)}, \quad i = 1, \ldots, n, \tag{2.2}$$

where $\mathbf{z}_i$ is a $q \times 1$ vector of covariates observed on firm i (possibly equal to $\mathbf{x}_i$ in (2.1)), $(U_1^{(2)}, \ldots, U_n^{(2)})^T$ are random errors following a n-variate normal distribution with mean $E(\boldsymbol{U}^{(2)}) = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}^{(2)} = \sigma^{(2)}\boldsymbol{I}_n$. As will be discussed later, we do not require the independence of the error terms in (2.1) and (2.2); differently from the traditional ML estimator, our method is robust and produces valid inferential results in the presence of dependent outcomes, for example due to serial correlation.

Let $\boldsymbol{\theta}$ be the overall parameter vector including parameters $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, $\sigma^{(1)}$ and $\sigma^{(2)}$. Based on (2.1) and (2.2), we can write the univariate zero-inflated distribution of the observed response Y_i as follows:

$$P(Y_i = l; \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) = \begin{cases} P(S_i = 0; \boldsymbol{\beta}^{(1)}) + P(S_i = 1; \boldsymbol{\beta}^{(1)}) P(Y_i = 0 | S_i = 1, \boldsymbol{\beta}^{(2)}, \boldsymbol{\gamma}), & l = 0, \\ P(S_i = 1; \boldsymbol{\beta}^{(1)}) P(Y_i = l | S_i = 1; \boldsymbol{\beta}^{(2)}, \boldsymbol{\gamma}), & l \neq 0, \end{cases} \tag{2.3}$$

where $P(S_i = 0; \boldsymbol{\beta}^{(1)}) = \Phi(-\mathbf{x}_i^\top \beta^{(1)})$, $P(S_i = 1; \boldsymbol{\beta}^{(1)}) = 1 - P(S_i = 0; \boldsymbol{\beta}^{(1)})$, with $\Phi(\cdot; v)$ denoting the cumulative distribution function of a univariate normal random variable with variance v. The conditional probability of Y_i given $S_i = 1$ is then written explicitly as

$$P(Y_i = l | S_i = 1; \boldsymbol{\beta}^{(2)}, \boldsymbol{\gamma}) = \Phi(\gamma_{l+1} - \mathbf{z}_i^\top \beta^{(2)}) - \Phi(\gamma_l - \mathbf{z}_i^\top \beta^{(2)}), \quad l = 0, \ldots, K,$$

where $\gamma_{K+1} = \infty$ and $\gamma_0 = -\infty$.

Robust inference by exponential tilting

Let $\pi = (\pi_1, \ldots, \pi_n)^\top$ be a vector of multinomial probabilities on n points. Given n observations $y_1, \ldots, y_n$, we define the tilted log-likelihood function by

$$\ell_{\boldsymbol{\pi}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \pi_i \log P(Y_i = y_i; \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}), \tag{2.4}$$

where $P(Y_i = y_i; \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})$ is defined in (2.3). The π_i's are regarded as sampling probabilities applied to the n observations under non-parametric bootstrap re-sampling Hall (2000); Genton and Hall (2015). We propose weighting by π_i the contribution from Y_i to the log-likelihood, and choose π_i to be small when Y_i is relatively unlikely to come from a population with the ZIIR model given in Section II. To this end, we compute tilting probabilities $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}(\boldsymbol{\theta})$ solving the following program:

$$\max_{\boldsymbol{\pi}} \ell_{\boldsymbol{\pi}}(\boldsymbol{\theta}), \quad \text{s.t.: } D_{KL}(\boldsymbol{\pi}; \boldsymbol{\pi}^{unif}) = \delta, \ \sum_{i=1}^{n} \pi_i = 1, \tag{2.5}$$

where

$$D_{KL}(\boldsymbol{\pi}; \boldsymbol{\pi}^{unif}) = \sum_{i=1}^{n} \pi_i \log(n\pi_i) \tag{2.6}$$

represents the Kullback-Leibler divergence between the candidate sampling probability $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$ and the uniform weights $\boldsymbol{\pi}^{unif} = (1/n, \ldots, 1/n)^\top$. The resulting weights may be interpreted as obtained by tilting a uniform prior, on the sample, so as to move away a given distance δ, from the uniform distribution. Such a movement occurs in the direction that gives most emphasis to data that enjoy larger likelihood.

By the Lagrange multiplier method, it is easy to check that the solution of the optimization problem (2.5) has the folliwng explicit solution for the bootstrap weights

$$\hat{\pi}_i^{(\alpha)}(\boldsymbol{\theta}) = \frac{P(Y_i = y_i; \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})^\alpha}{\sum_{j=1}^{n} P(Y_j = y_j; \mathbf{x}_j, \mathbf{z}_j, \boldsymbol{\theta})^\alpha}, \quad i = 1, \ldots, n, \tag{2.7}$$

where $\alpha \geq 0$ is a constant that may be found, given δ, by solving the equation $D_{KL}(\hat{\boldsymbol{\pi}}^{(\alpha)}) = \delta$. Due to this correspondence between δ and α, we avoid solving this last equation and focus directly on α as a tuning constant for the tilting method.

Given a desired level of tilting $\alpha \geq 0$, the tilted estimator $\hat{\boldsymbol{\theta}}^{(\alpha)}$ is found by replacing $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}^{(\alpha)}(\boldsymbol{\theta})$ in (2.4) and then maximizing

$$\ell_{\hat{\boldsymbol{\pi}}^{(\alpha)}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \hat{\pi}_i^{(\alpha)}(\boldsymbol{\theta}) \times \log P(Y_i = y_i; \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) \tag{2.8}$$

with respect to $\boldsymbol{\theta}$. Maximization of the above objective function is carried out numerically using a standard Newton-type algorithm. When $\alpha = 0$, we have $\hat{\pi}_i^{(\alpha)}(\boldsymbol{\theta}) = 1$ for all $i = 1, \ldots, n$, which corresponds to standard ML estimation. In this case, no adjustment of the data distribution through tilting occurs and the final estimates might be affected by a few very unusual observations corresponding to poor data likelihood. For $\alpha > 0$, the impact of the contribution from the ith unit on the overall log-likelihood is moderated by the weight $\pi_i^{(\alpha)}(\boldsymbol{\theta})$; specifically, observations with larger likelihood will contribute substantially to (2.8), whereas observations incompatible with the assumed model will receive a relatively low weight.

Under appropriate regularity conditions, standard arguments from M-estimation theory may be applied to derive the asymptotic normal distribution of the tilted estimator $\hat{\boldsymbol{\theta}}^{(\alpha)}$ Choi et al. (2000). Its covariance matrix takes the usual sandwich form

$$\mathbf{V}^{(\alpha)}(\boldsymbol{\theta}) = \mathbf{J}^{(\alpha)}(\boldsymbol{\theta})^{-1} \mathbf{K}^{(\alpha)}(\boldsymbol{\theta}) \mathbf{J}^{(\alpha)}(\boldsymbol{\theta})^{-1}, \tag{2.9}$$

where $\mathbf{J}^{(\alpha)}(\boldsymbol{\theta}) = -E\{\nabla^2 \ell_{\hat{\pi}^{(\alpha)}}(\boldsymbol{\theta})\}$ and $\mathbf{K}^{(\alpha)}(\boldsymbol{\theta}) = Var\{\nabla \ell_{\hat{\pi}^{(\alpha)}}(\boldsymbol{\theta})\}$ are the sensitivity and variability matrices, respectively, with ∇ and ∇^2 denoting, respectively, the gradient and hessian operators with respect to the parameter vector $\boldsymbol{\theta}$. An estimate $\hat{\mathbf{V}}_{boot}^{(\alpha)}$ of $\mathbf{V}^{(\alpha)}$ is computed as

$$\hat{\mathbf{V}}_{boot}^{(\alpha)} = \frac{n}{R-1} \sum_{r=1}^{R} \left(\hat{\boldsymbol{\theta}}_r^{(\alpha)} - \bar{\hat{\boldsymbol{\theta}}}^{(\alpha)} \right) \left(\hat{\boldsymbol{\theta}}_r^{(\alpha)} - \bar{\hat{\boldsymbol{\theta}}}^{(\alpha)} \right)^\top, \tag{2.10}$$

where $\hat{\boldsymbol{\theta}}_1^{(\alpha)}, \ldots, \hat{\boldsymbol{\theta}}_R^{(\alpha)}$ are estimates of $\boldsymbol{\theta}$ for given α computed via standard non-parametric bootstrap techniques and $\bar{\hat{\boldsymbol{\theta}}}^{(\alpha)} = R^{-1} \sum_{r=1}^{R} \hat{\boldsymbol{\theta}}_r^{(\alpha)}$ denotes their average.

Selection of the tilting constant α

The choice of the appropriate amount of tilting may be performed using a bootstrap approach. In particular, we choose α by minimizing the following bootstrap estimate of the mean squared error (MSE) of $\hat{\boldsymbol{\theta}}^{(\alpha)}$:

$$\widehat{MSE}_{boot}(\alpha) = \left\| \hat{\boldsymbol{\theta}}^{(\alpha)} - \bar{\hat{\boldsymbol{\theta}}}^{(\alpha)} \right\|_2^2 + n^{-1}\mathrm{Tr}\left(\hat{\mathbf{V}}_{boot}^{(\alpha)}\right), \tag{2.11}$$

where $\mathrm{Tr}(\mathbf{A})$ denotes the trace of the matrix $\mathbf{A}$, $\bar{\hat{\boldsymbol{\theta}}}^{(\alpha)} = R^{-1}\sum_{r=1}^{R}\hat{\boldsymbol{\theta}}_r^{(\alpha)}$ is the average of R bootstrapped estimates and $\hat{\mathbf{V}}_{boot}^{(\alpha)}$ is their sample covariance matrix given in (2.10). When contamination is absent and all the observations are compatible with the assumed data generating process, no tilting is required and the best choice is $\alpha = 0$ (corresponding to the ML estimator) since this option guarantees optimal MSE in large samples. In the presence of deviations of the data distribution from the assumed model, estimation accuracy is improved by letting $\alpha > 0$. We found in numerical experiments that values of α greater than 1 correspond to overly large bootstrap MSE estimates, whereas the best performance is recorded when $0 < \alpha \leq 1$.

III. Monte Carlo experiments

To assess the accuracy of the proposed method, Monte Carlo samples of size n are generated to include $(1-\epsilon)\times n$ clean data from the ZIIR model in Section II and $\epsilon \times n$ outliers. We consider various specifications of the contamination level $0 \leq \epsilon \leq 0.5$ and tilting parameter α. Results are then compared to the ML approach, corresponding to $\alpha = 0$. The simulation setup is based on parameter values equal to $\boldsymbol{\beta}^{(1)} = (-0.1, 0.2)^\top$, $\boldsymbol{\beta}^{(2)} = (2,2)^\top$, $\boldsymbol{\gamma} = (-5,0,5)^\top$, and $\sigma^{(1)} = \sigma^{(2)} = 1$. The two fixed covariates in each vector $\mathbf{x}_i$ are generated as independent draws from a standard bivariate normal distribution. In the case of no contamination with $\varepsilon = 0$, such a setting produces an inflated zero class frequency and sizable frequencies for the remaining positive classes. When $\varepsilon > 0$, outliers are generated according to the three experimental settings detailed below.

(i) *Experiment 1: Contaminated hurdle probit.* The contamination occurs in the latent component of the first-stage probit model (2.1). Specifically, a contaminated sample is obtained by generating $\epsilon \times n$ realizations of S^* from the rescaled and shifted Gaussian distribution $N(\mu_\epsilon^{(1)} + \mathbf{x}_i^\top\boldsymbol{\beta}^{(1)}, \sigma^{(1)}/10)$, where $\mu_\epsilon^{(1)}$ is a location shift in the latent component of the probit. For illustration purposes, we present results for $\mu_\epsilon^{(1)} = -5$, which mimics a relatively high proportion of zeros.

(ii) *Experiment 2: Contaminated latent regression model.* The contamination occurs in the latent component of the second-stage regression model in (2.2). Specifically, a contaminated sample is obtained by generating $\epsilon \times n$ realizations of Y^* from the rescaled and shifted Gaussian distribution $N(5 + \mathbf{x}_i^\top\boldsymbol{\beta}^{(2)}, \sigma^{(2)}/2)$, where $\mu_\epsilon^{(2)}$ is a location shift in the latent component of the probit. For illustration purposes, we present results for $\mu_\epsilon^{(2)} = 5$ to obtain asymmetric outliers affecting the right tail of the error distribution.

(iii) *Experiment 3: Correlated latent errors.* We draw $\epsilon \times n$ observations of Y from the distribution specified in (2.3) with autocorrelated error terms at both modeling stages. In more detail, we set $(U_1^{(s)}, \ldots, U_n^{(s)}) \sim N_n(\mathbf{0}, \boldsymbol{\Sigma})$ for $s = 1, 2$, where $\boldsymbol{\Sigma}$ is a

correlation matrix with unit diagonal elements and off-diagonal elements equal to $0 < \rho < 1$. For illustration purposes, we consider $\rho = 0.25, 0.75$.

The performance of the tilted estimator is evaluated in terms of the MSE computed over $B = 1000$ Monte Carlo runs

$$\widehat{MSE}_{MC} = \frac{1}{B}\sum_{b=1}^{B}\left\|\hat{\boldsymbol{\theta}}_b^{(\alpha)} - \boldsymbol{\theta}_0\right\|^2, \tag{3.12}$$

where $\hat{\boldsymbol{\theta}}_b^{(\alpha)}$, $b = 1, \ldots, B$ is the vector of parameter estimates obtained at each run for a given tuning constant α, and $\boldsymbol{\theta}_0$ is the true vector of coefficients.

Tables 1 - 3 show Monte Carlo MSEs with simulation standard errors reported in parentheses. As expected, the value $\alpha = 0$ corresponding to no tilting fails to ensure robustness in all the considered settings, with the MSE typically increasing with the size of contamination level ϵ. On the other hand, values of the tilting parameter $\alpha > 0$ afford greater robustness to outliers. As the contamination level grows, the optimal α minimizing the MSE also tends to increase with values between $\alpha = 0.3$ and $\alpha = 0.7$ working reasonably well across several experimental scenarios.

TABLE 1
Monte Carlo MSE estimates of the model described in Experiment 1.

α	ϵ						
	0	0.05	0.10	0.20	0.30	0.40	0.50
0	0.632	0.487	0.437	0.433	0.678	1.282	3.814
	(0.092)	(0.083)	(0.082)	(0.079)	(0.103)	(0.170)	(0.335)
0.01	0.589	0.477	0.425	0.411	0.640	1.222	3.397
	(0.089)	(0.083)	(0.082)	(0.078)	(0.099)	(0.166)	(0.285)
0.05	0.462	0.386	0.345	0.324	0.517	0.876	2.525
	(0.074)	(0.073)	(0.073)	(0.068)	(0.084)	(0.116)	(0.237)
0.10	0.406	0.326	0.284	0.254	0.400	0.634	1.721
	(0.071)	(0.067)	(0.067)	(0.060)	(0.071)	(0.088)	(0.175)
0.30	0.101	0.081	0.067	0.085	0.216	0.294	0.527
	(0.036)	(0.036)	(0.037)	(0.041)	(0.060)	(0.054)	(0.070)
0.50	0.067	0.053	0.045	0.041	0.055	0.093	0.211
	(0.029)	(0.029)	(0.029)	(0.029)	(0.032)	(0.036)	(0.050)
0.70	0.054	0.048	0.046	0.043	0.046	0.062	0.099
	(0.025)	(0.026)	(0.027)	(0.028)	(0.031)	(0.034)	(0.039)
1	0.070	0.064	0.063	0.058	0.055	0.061	0.075
	(0.026)	(0.027)	(0.029)	(0.029)	(0.032)	(0.036)	(0.040)

Notes: Each simulated sample, based on 1000 runs of the model, includes a fraction ϵ of observations obtained by contaminating the latent component of the first-stage probit model. Simulation standard errors are reported in parentheses.

TABLE 2
Monte Carlo MSE estimates of the model described in Experiment 2.

α	ϵ						
	0	0.05	0.10	0.20	0.30	0.40	0.50
0	0.632	0.496	0.428	0.413	0.411	0.701	0.731
	(0.092)	(0.100)	(0.093)	(0.108)	(0.118)	(0.126)	(0.122)
0.01	0.589	0.465	0.406	0.407	0.425	0.770	0.821
	(0.089)	(0.098)	(0.090)	(0.105)	(0.115)	(0.120)	(0.116)
0.05	0.462	0.397	0.374	0.441	0.560	1.091	1.204
	(0.074)	(0.093)	(0.084)	(0.096)	(0.105)	(0.101)	(0.098)
0.10	0.406	0.394	0.422	0.592	0.823	1.521	1.674
	(0.071)	(0.092)	(0.083)	(0.092)	(0.094)	(0.082)	(0.078)
0.30	0.101	0.771	0.978	1.452	2.026	2.980	3.131
	(0.036)	(0.095)	(0.077)	(0.067)	(0.059)	(0.044)	(0.044)
0.50	0.067	1.390	1.677	2.289	2.959	3.907	4.011
	(0.029)	(0.101)	(0.075)	(0.057)	(0.045)	(0.030)	(0.031)
0.70	0.054	1.940	2.259	2.926	3.611	4.512	4.582
	(0.025)	(0.101)	(0.071)	(0.048)	(0.036)	(0.023)	(0.024)
1	0.070	2.580	2.919	3.611	4.280	5.104	5.143
	(0.026)	(0.095)	(0.063)	(0.039)	(0.028)	(0.017)	(0.018)

Notes: Each simulated sample, based on 1000 runs of the model, includes a fraction ϵ of observations obtained by contaminating the latent component of the second-stage probit model. Simulation standard errors are reported in parentheses.

Table 3 reports simulation outputs referred to Experiment 3, where latent errors exhibit positive autocorrelation ρ in both first-stage hurdle probit and second-stage ordered regression models. In order to investigate the effect of the correlation on estimation accuracy, we show results obtained when the correlation is mild ($\rho = 0.25$) or strong ($\rho = 0.75$). Note that the bias given by the increased correlation can be severe, with the MSE inflating dramatically under strong contamination with $\epsilon = 0.5$. Adjusting the amount of tilting by raising α appears to be quite effective in mitigating the undesired effects of the error autocorrelation.

IV. An application to cyber security survey data

Cyber security attacks have long been acknowledged as a severe threat to companies and the entire economy (Kelly, 1999; CEA, 2018). In recent years, an increasing amount of funds is spent worldwide on cyber security programs and awareness campaigns. Despite this increasing interest, there is a substantial lack of applied studies to help identify the factors associated with the cost of cyber breaches. The analysis in this paper contributes to our understanding of the relationship between investments in cyber defences and costs occurred from cyber attacks at the firm level. Empirical results in this area provide useful information to facilitate the development of well-targeted economic theory and managerial policies.

TABLE 3
Monte Carlo MSE estimates of the model described in Experiment 3.

α	ϵ						
	0	0.05	0.10	0.20	0.30	0.40	0.50
				$\rho = 0.25$			
0	0.624	0.622	0.619	0.624	0.637	0.643	0.673
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)
0.01	0.590	0.591	0.588	0.591	0.609	0.614	0.646
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.010)
0.05	0.484	0.484	0.482	0.483	0.498	0.506	0.534
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
0.10	0.410	0.410	0.401	0.397	0.408	0.408	0.431
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
0.30	0.104	0.105	0.107	0.110	0.119	0.131	0.148
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
0.50	0.075	0.076	0.075	0.075	0.077	0.081	0.086
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
0.70	0.056	0.057	0.057	0.057	0.061	0.065	0.068
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
1	0.073	0.074	0.074	0.075	0.077	0.081	0.083
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
				$\rho = 0.75$			
0	0.624	0.621	0.625	0.633	0.694	0.768	1.023
	(0.009)	(0.009)	(0.009)	(0.010)	(0.012)	(0.014)	(0.023)
0.01	0.590	0.590	0.594	0.605	0.670	0.738	0.959
	(0.008)	(0.008)	(0.009)	(0.010)	(0.012)	(0.014)	(0.020)
0.05	0.484	0.483	0.487	0.498	0.559	0.618	0.807
	(0.007)	(0.007)	(0.007)	(0.008)	(0.010)	(0.011)	(0.016)
0.10	0.410	0.408	0.395	0.400	0.447	0.498	0.637
	(0.006)	(0.006)	(0.006)	(0.007)	(0.008)	(0.010)	(0.013)
0.30	0.104	0.106	0.111	0.123	0.165	0.211	0.275
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.005)	(0.007)
0.50	0.075	0.075	0.077	0.081	0.096	0.110	0.137
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)
0.70	0.056	0.057	0.058	0.062	0.073	0.085	0.099
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)
1	0.073	0.074	0.075	0.078	0.087	0.094	0.104
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)	(0.004)

Notes: Monte Carlo MSE estimates based on 1000 samples of size $n = 1000$. Each simulated sample includes a fraction ϵ of observations with correlated errors in the first-stage probit and second-stage ordered probit models (correlation $\rho = 0.25, 0.75$). Simulation standard errors are reported in parenthesis.

Data

The Cyber Security Breaches Survey (CSBS) is an annual study of businesses and charities in the United Kingdom. It represents one well-established official data source on cyber

security collected recurrently by a government body. The UK government has been very active in monitoring risks from cyber attacks; in this regard, the CSBS influences how the government shapes future policy, allows organizations to compare their cyber security with others and demonstrates the trends in this rapidly evolving area. Here we examine a data set obtained by combining two recent surveys taken between October and February of 2017/2018 and 2018/2019. Interviewed companies and charities were asked about their approach to cyber security and any breaches over the 12 months preceding the interview. In our analysis, we focus on for-profit companies. The resulting data are not panel data since the two years may not contain measurements on the same firms and firms are assigned different anonymous labels in the two years. The presence of repeated measurements on the same firms violates the standard assumption of independence for the errors in the standard ZIIR model described in Equations (2.1) and (2.2).

In the remainder of this section we overview the main features of our data set; for a complete description of the survey methodology and variables, see the technical annex provided on the CSBS website[2]. The main response of interest is the cost related to all breaches experienced in the 12 months preceding the interview. Information on the cost is collected through the question *"Approximately how much, if anything, do you think the cyber security breaches or attacks you have experienced in the last 12 months have cost your organization financially?"*. In order to prevent the possibility of individual organizations being identified, only banded costs were made publicly available instead of the exact cost figures. Table 4 shows the observed frequencies for each cost band. About 23% of the companies interviewed declared costs in the first interval class between 0 and 500 GBP. This class includes companies that had exactly zero costs mixed with those sustaining small yet non-zero costs.

TABLE 4
Observed frequency (N) and percent relative frequency (%) for the cost intervals in the 2018, 2019 surveys and in the aggregated data.

Cost intervals (1000 GBP)	2018		2019		Combined	
	N	%	N	%	N	%
$[0, 0.5)$	36	23.8	29	21.5	65	22.7
$[0.5, 1)$	19	12.6	19	14.1	38	13.3
$[1, 5)$	38	25.2	42	31.1	80	28.0
$[5, 10)$	19	12.6	21	15.6	40	14.0
$[10, 20)$	18	11.9	7	5.2	25	8.7
$[20, 50)$	11	7.3	11	8.1	22	7.7
$[50, 100)$	3	2.0	3	2.2	6	2.1
$[100, \infty)$	7	4.6	3	2.2	10	3.5
Total	151	100.0	135	100.0	286	100.0

In our empirical study, we focus on the impact of various independent variables on the cost response. Table 5 lists the main predictors considered, with their description and summary statistics. To assess whether investments affect the costs associated with cyber

[2]https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2019

security over time, the main predictor of interest is the banded variable Invest, representing the GBP amount invested in cyber security. Other relevant predictors are the number of breaches (Numbb), and the presence of a cyber security incident management process (Incid). Industry type and company's size are controlled for by dummies for 12 industrial sectors (UK Standard Industrial Classification) and the sales turnover in million GBP (Sales), respectively. Observations on the above variables are collected through the questions: *"In the financial year just gone, approximately how much, if anything, did you invest in cyber security?"*; *"Approximately, how many breaches or attacks have you experienced in total across the last 12 months?"*; *"Do you have any formal cyber security incident management process, or not?"*. The time predictor is coded as a numerical variable taking values 1 and 2 for 2018 and 2019, respectively.

TABLE 5
Main predictors

Predictor	Description	2018		2019	
		Mean	SE	Mean	SE
Incid	Incident management process available? (%)	34.4	3.9	46.7	4.3
Sales	Turnover (1,000,000 GBP)	16.1	1.5	14.6	1.6
Invest	Investment in the last 12 months (1,000 GBP)	137.2	44.1	118.7	38.7
Numbb	Number of attacks in the last 12 months	139	44	1,626	1,043
Admin	Administration or real estate sector (%)	7.3	2.1	11.1	2.7
Constr	Construction sector (%)	13.9	2.8	8.1	2.3
Educa	Education sector (%)	6.0	1.9	5.9	2.0
Entert	Entertainment, service or membership sector (%)	5.3	1.8	4.4	1.8
FinIns	Finance or insurance sector (%)	9.3	2.4	6.7	2.1
FoodHos	Food or hospitality sector (%)	6.0	1.9	2.2	1.3
Health	Health, social care or social work sector (%)	4.0	1.6	5.2	1.9
InfoCom	Information or communication sector (%)	8.6	2.3	7.4	2.2
SciPro	Professional, scientific or technical sector (%)	13.9	2.8	12.6	2.8
Retail	Retail or wholesale sector (%)	13.9	2.8	14.8	3.0
Transp	Transport or storage sector (%)	4.6	1.7	9.6	2.5
Util	Utilities or production sector (%)	7.3	2.1	11.9	2.8

Notes: Sample means and sample proportions (Mean) computed for banded and dummy variables, respectively, with corresponding standard errors (SE). For banded variables, the mid-point of each interval class is considered.

Results

Parameter estimates for the ZIIR models corresponding to the optimal tilting parameter $\alpha = 0.74$ are reported in Table 6; the tilting constant was selected via the bootstrap MSE minimization described in Section II. ML estimates without tilting (corresponding to $\alpha = 0$) are also shown for comparison. The response variable in our model is the cost intervals mid-point transformed on the logarithmic scale. In the first stage equation, we included the industrial sectors as predictors for the probability of non-zero costs; in the second stage equation, we considered all the remaining predictors listed in Table 5.

To assess the potential change of investment and its effect on costs, we included the interaction between time and investment (Time×Invest). We consider the logarithm of the interval mid-point for all banded predictors. In order to enable comparison, we report the results of four alternative models in Table 7: probit and logit models, linear regression estimated by ordinary least squares (OLS), and interval regression (IR) estimated by ML. For the probit and logit models, the response variable is coded as 0 for the first class containing 0 (representing costs between 0 and 500 GBP) and 1 for the other classes (costs greater than 500 GBP). For linear regression, the response is coded by taking the mid-point of each interval class and then transforming on the logarithmic scale.

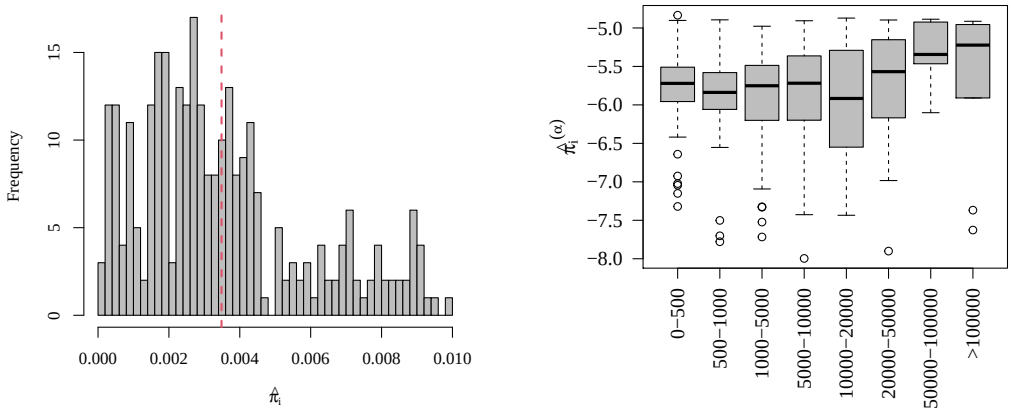

Figure 1. Left: Distribution of the tilted sampling weights $\hat{\pi}_i^{(\alpha)}$ with $\alpha = 0.74$ for the CSBS survey data; the vertical line represents the original sampling weight $1/n$ corresponding to $\alpha = 0$. Right: Box-plots showing the distribution of the tilted sampling weights $\hat{\pi}_i^{(\alpha)}$ (on the log-scale) grouped by cost bands.

Figure 1 (left) shows the distribution of the tilted sampling weights $\hat{\pi}_i^{(\alpha)}$ with $\alpha = 0.74$ with the vertical line representing the original sampling weight $1/n$ corresponding to $\alpha = 0$. Figure 1 (right) shows the distribution of the tilted sampling weights on the log-scale grouped by cost band. Whereas a fair proportion of the data receive relatively large weights (larger than $1/n$), the sample also contains many observations which appear to be inconsistent with the ZIIR model specifications. The inconsistent observations are more likely to occur for moderate or lower costs. This may be partly due to the the presence of a substantial number of statistical units surveyed in both sampled years.

The estimates of the first-stage equation of the tilted version of the model reveals some degree of heterogeneity among the industrial sectors. The Administration or Real Estate sector exhibits the highest coefficient (5.67), meaning that firms in that sector are more likely to sustain losses from cyber attacks. On the other hand, the Health sector has the lowest coefficient (1.08) and is the least likely affected. The probit and logit models in Table 7 confirm the estimated sizes – including the relevance of Administration or Real Estate sector.

TABLE 6
Estimates for the zero-inflated interval regression model with and without exponential tilting.

	ZIIR ($\alpha = 0$)		Tilted ZIIR ($\alpha = 0.74$)	
	$\hat{\beta}_j^{(1)}$ (SE)	$\hat{\beta}_j^{(2)}$ (SE)	$\hat{\beta}_j^{(1)}$ (SE)	$\hat{\beta}_j^{(2)}$ (SE)
Const	–	$-6.03(1.53)\star$	–	$-4.37(1.05)\star$
Incid	–	$0.09(0.22)$	–	$-0.27(0.18)$
Sales	–	$0.19(0.07)\star$	–	$0.16(0.05)\star$
Invest	–	$0.50(0.16)\star$	–	$0.38(0.10)\star$
Time	–	$1.28(0.96)$	–	$1.28(0.71)\star$
Invest×Time	–	$-0.16(0.10)$	–	$-0.13(0.07)\star$
Numbb	–	$0.07(0.04)$	–	$0.05(0.04)$
Admin	$11.18(24.57)$	–	$5.67(1.96)\star$	–
Constr	$1.06(1.05)$	–	$1.28(0.56)\star$	–
Educa	$1.01(2.01)$	–	$1.29(1.59)$	–
Entert	$1.54(2.61)$	–	$1.60(2.20)$	–
FinIns	$1.35(2.25)$	–	$1.27(1.50)$	–
FoodHos	$5.21(6.84)$	–	$1.28(3.55)$	–
Health	$-0.02(0.58)$	–	$1.08(0.27)\star$	–
InfoCom	$1.02(1.56)$	–	$1.31(0.72)$	–
SciPro	$1.12(6.67)$	–	$1.22(0.27)\star$	–
Retail	$0.99(0.47)\star$	–	$1.37(0.15)\star$	–
Transp	$0.80(1.50)$	–	$1.14(0.51)\star$	–
Util	$0.69(0.77)$	–	$1.11(0.18)\star$	–
$\widehat{MSE}_{boot}(\alpha)$	735.43		33.78	

Notes: In parentheses, we report the bootstrap standard error obtained based on 2000 bootstrap re-samples. Coefficients different from zero at the 0.05 level of significance are marked by "⋆". The last line reports the mean squared error statistics $\widehat{MSE}_{boot}(\alpha)$ given in Section II.

The estimates for the second-stage interval regression model highlight a significant interaction between Investments and Time. The estimated effect of investments on the amount of cost changes from $0.38 - 0.13 \times 1 = 0.25$ in 2018 to $0.38 - 0.13 \times 2 = 0.12$ in 2019, showing a nearly twofold reduction in the effect of investment within one year. The standard OLS and IR models confirm this result by showing significant negative interactions. These findings support the hypothesis that the investments are effective in reducing the cost amount of a cyber attack. The estimates for the second-stage model also indicate a positive significant effect on the response of the variable Sales, which is found statistically significant also by OLS model in Table 7. The estimated effect of firm size (measured here in terms of sales turnover) is in line with the results obtained by Romanosky (2016) and Aldasoro et al. (2020). The estimated coefficient for the Investment main effect is positive and significant. This is probably due to endogeneity, as also noted by Gandal et al. (2020) and by Woods and Böhme (2021).

Figure 2 compares the observed frequency for the cost bands with the estimated frequencies for $\alpha = 0$ and $\alpha = 0.74$. Both choices give a reasonable fit with Pearson chi-

TABLE 7
Estimates from four alternative regression models

	Probit	Logit	OLS	IR
	$\hat{\beta}_j$ (SE)	$\hat{\beta}_j$ (SE)	$\hat{\beta}_j$ (SE)	$\hat{\beta}_j$ (SE)
Const	–	–	−6.13(1.38)⋆	−76.02(25.78)⋆
Sales	–	–	0.15(0.06)⋆	1.22(1.11)
Invest	–	–	0.55(0.13)⋆	8.03(1.62)⋆
Time	–	–	1.39(0.78)	26.11(13.56)
Invest×Time	–	–	−0.17(0.08)⋆	−3.32(1.18)⋆
Numbb	–	–	0.07(0.04)	0.26(0.77)
Incid	–	–	−0.14(0.22)	2.44(3.76)
Admin	1.20(0.32)⋆	2.04(0.61)⋆	–	–
Constr	0.78(0.25)⋆	1.27(0.43)⋆	–	–
Educa	0.72(0.33)⋆	1.18(0.57)⋆	–	–
Entert	1.07(0.41)⋆	1.79(0.76)⋆	–	–
FinIns	1.12(0.33)⋆	1.90(0.62)⋆	–	–
FoodHosp	0.97(0.43)⋆	1.61(0.77)⋆	–	–
Health	−0.10(0.35)	−0.15(0.56)	–	–
InfoCom	0.81(0.29)⋆	1.33(0.50)⋆	–	–
SciPro	0.63(0.22)⋆	1.03(0.37)⋆	–	–
Retail	0.86(0.22)⋆	1.42(0.39)⋆	–	–
Transp	0.67(0.30)⋆	1.10(0.52)⋆	–	–
Util	0.43(0.25)	0.69(0.41)	–	–

Notes: IR: ordinal probit interval regression. Standard errors are reported in parentheses. Coefficients different from zero at the 0.05 level of significance are marked by "⋆".

square statistics comparing the observed and expected distributions of 3.58 and 5.40 for $\alpha = 0$ and $\alpha = 0.74$, respectively (corresponding to p-values of 0.82 and 0.61 based on a chi-square distribution with seven degrees of freedom). However, the estimated MSE improves dramatically, dropping from 735.43 to 33.78 when α is increased from 0 to 0.74. The superior accuracy of the tilted estimates is also confirmed by the bootstrap standard errors of the tilted estimator, which are clearly smaller than those obtained with $\alpha = 0$. Overall our empirical results confirm the findings from the Monte Carlo simulations presented in Section III.

V. Conclusions

We consider estimation by exponential tilting of the ZIIR model that allows for zero and non-zero observations to be generated by two different behavioural regimes. The proposed method consists of tilting the empirical distribution of the data so to emphasize observations with greater likelihood. The tilting procedure is a robust generalization of the ML method: when the tilting parameter equals $\alpha = 0$, observations receive equal weights; when $\alpha > 0$, the resulting estimator is shown to mitigate issues related to the presence of observations inconsistent with the ZIIR model, and enables one to detect the presence of potential outliers in survey data.

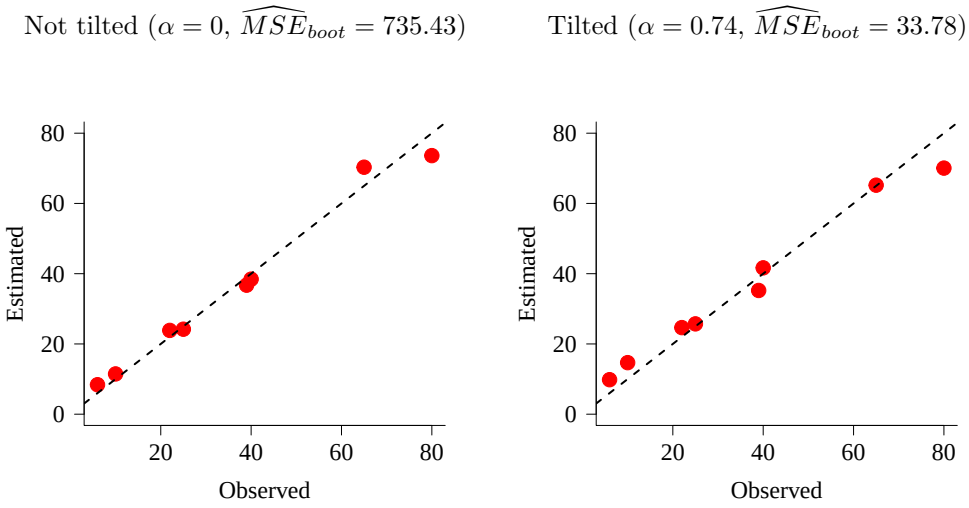

Figure 2. Observed versus estimated frequencies for the response cost intervals without tilting ($\alpha = 0$) and with tilting ($\alpha = 0.74$). Computation of the estimated frequencies and selection of the optimal tilting parameter are described in Section II.

Our empirical analyses highlight several advantages of the tilted estimator for the ZIIR model. Monte Carlo experiments show that tilting improves estimation accuracy in the presence of data deviating from the assumed ZIIR model under different scenarios concerning the actual data generating process. Particularly, when a fraction of observations has correlated errors in the latent equations, the ZIIR model estimated via tilting clearly outperforms the non-tilted ML estimator. This aspect is relevant for the analysis of the CSBS data set, where correlation among units is likely to occur since each company can be surveyed at the two time points. The increased accuracy resulting from tilting is also reflected in our empirical studies by the smaller size of the standard errors for the parameter estimates and of the bootstrap MSE with respect to those obtained via ML estimation.

The tilted ZIIR model is applied to the CSBS data to evaluate the impact of investments in cyber security on costs suffered by companies following cyber attacks and breaches. Despite the increasing interest of the private and public sectors in the economics of cyber security, which mirrors the increased number of cyber attacks, the research on the determinants of the costs from cyber breaches is still in its infancy. Our study provides new evidence supporting the effectiveness of investments in cyber security. Robust estimates obtained via tilting clearly show an effect of the investments in reducing the amount of the loss from a cyber breach over time. The firm size is another significant determinant of the costs.

Although our study provides important elements to better understand the costs arising from cyber breaches, further analyses will be useful, possibly leading to modifications of the basic ZIIR model for cyber security survey data. One advantage of the exponential tilting procedure is the ability to produce ranks of the observations with respect to their degree of compatibility with the nominal ZIIR model based on the tilted weight. A more detailed examination of the observations with relatively low weight would then lead

to useful modifications of the theoretical model assumptions. For example, a suitable correlation structure for the errors in the model Equations (2.1) and (2.2) should be explored, possibly in relation with firm demographics. Moreover, the second stage of the model could further specify the type of attack (or the type of investment) for investigating whether certain attacks are more damaging (or some investments are more effective in mitigating the losses). Finally, the role of sector heterogeneity also requires further investigations: for instance, a pair-wise analysis of firms could provide insights on the relative exposure to losses in each industrial sector.

References

Aldasoro, I., L. Gambacorta, P. Giudici, and T. Leach (2020). The drivers of cyber risk. *BIS Working Papers No 865, May 2020* (865).

Bagozzi, B. E., D. W. Hill Jr., W. H. Moore, and B. Mukherjee (2015). Modeling Two Types of Peace: The Zero-inflated Ordered Probit (ZiOP) Model in Conflict Research. *Journal of Conflict Resolution 59*(4), 728–752.

Biancotti, C. (2017). The price of cyber (in)security: evidence from the italian private sector. *Questioni di Economia e Finanza (Occasional Paper)* (407).

Brown, S., A. Duncan, M. N. Harris, J. Roberts, and K. Taylor (2015). A zero-inflated regression model for grouped data. *Oxford Bulletin of Economics and Statistics 77*(6), 822–831.

Camponovo, L. and T. Otsu (2012). Breakdown point theory for implied probability bootstrap. *The Econometrics Journal 15*(1), 32–55.

Choi, E., P. Hall, and B. Presnell (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika 87*(2), 453–465.

Council of Economic Advisers (CEA) (2018). *The Costs of Maliciuos Cyber Activity to the U.S. Economy*. Executive Office of the President of the United States.

Critchley, F. and P. Marriott (2004). Data-informed influence analysis. *Biometrika 91*(1), 125–140.

Das, U. and K. Das (2018). Inference on zero inflated ordinal models with semiparametric link. *Computational Statistics and Data Analysis 128*, 104–115.

Downward, P., F. Lera-Lopez, and S. Rasciute (2011). The Zero-Inflated ordered probit approach to modelling sports participation. *Economic Modelling 28*, 2469–2477.

Eling, M. and J. Wirfs (2019). What are the actual costs of cyber risk events? *European Journal of Operational Research 272*(3), 1109–1119.

Ferrari, D. and C. Zheng (2016). Reliable inference for complex models by discriminative composite likelihood estimation. *Journal of Multivariate Analysis 144*, 68–80.

Gandal, N., M. Riordan, and S. Bublil (2020). A New Approach to Quantifying, Reducing and Insuring Cyber Risk: Preliminary Analysis and Proposal for Further Research. *CEPR Discussion Paper* (14461).

Genton, M. G. and P. Hall (2015). A tilting approach to ranking influence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1*(78), 77–97.

Gurmu, S. and G. A. Dagne (2012). Bayesian Approach to Zero-Inflated Bivariate Ordered Probit Regression Model, with an Application to Tobacco Use. *Journal of Probability and Statistics*.

Gurmu, S. and P. K. Trivedi (1996). Excess Zeros in Count Models for Recreational Trips. *Journal of Business and Economic Statistics 14*, 469–477.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics 56*(4), 1030–1039.

Hall, P. and B. Presnell (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 143–158.

Harris, M. N. and X. Zhao (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics 141*(2), 1073–1099.

Jiang, X., B. Huang, R. L. Zaretzki, S. Richards, X. Yan, and H. Zhang (2013). Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models. *Accident Analysis and Prevention 57*, 55–66.

Kelly, B. J. (1999). Preserve, Protect, and Defend. *Journal of Business Strategy 20*(5), 22–25.

Min, Y. and A. Agresti (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling 5*(1), 1–19.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics 33*, 341–365.

Pohlmeier, W. and V. Ulrich (1995). An Econometric Model of the Two-Part Decision-making Process in the Demand for Health Care. *Journal of Human Resources 30*, 339–361.

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity 2*(2), 121–135.

Tan, A. K. and S. T. Yen (2017). Cigarette consumption by individuals in Malaysia: a zero-inflated ordered probability approach. *Journal of Public Health 25*, 87–94.

Woods, D. W. and R. Böhme (2021). Systematization of Knowledge: Quantifying Cyber Risk. *In IEEE Symposium on Security and Privacy*.