Insurance: Mathematics and Economics 104 (2022) 222–242

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

www.elsevier.com/locate/ime

Unraveling heterogeneity in cyber risks using quantile regressions

Martin Eling [a], Kwangmin Jung [b,*], Jeungbo Shim [c]

[a] Institute of Insurance Economics, University of St. Gallen, Girtannerstrasse 6, 9010 St. Gallen, Switzerland
[b] Department of Industrial and Management Engineering, POSTECH (Pohang University of Science and Technology), Cheongam-ro 77, Nam-gu, Pohang, Gyeongbuk, 37673, South Korea
[c] Business School, University of Colorado Denver, 1475 Lawrence St, Denver, CO 80202, United States

ARTICLE INFO

Article history:
Available online 7 March 2022

JEL classification:
C21
C35
G2

Keywords:
Cyber risk
Quantile regression
Cyber cost estimation
Cyber-insurance pricing
Quantile premium principle

ABSTRACT

We consider quantile regressions for adequate cyber-insurance pricing across heterogenous policyholders and calculation of claims cost associated with data breach events. We show that the impact of a firm's revenue is stronger (weaker) in the lower (upper) quantile of the cost distribution. This result suggests that mispricing may occur if small and large firms are priced using the average effect estimated by the traditional least squares approach. Using a novel dataset, our study is the first to take firm-specific security information into account. We find that firms with weaker security levels than the industry average are more likely to be exposed to large-cost events. Regarding data breaches, small or mid-size loss events are related to higher cost per breached record. We compare the premiums of a quantile-based insurance pricing scheme with those of a two-part generalized linear model and the Tweedie model to explore the usefulness of the quantile-based model in addressing heterogeneous effects of firm size. Our findings provide useful implications for cyber insurers and policymakers who wish to assess the impacts of firm-specific factors in pricing insurance and to estimate the cost of claims.

1. Introduction

The increasing dependence on the information technology (IT) sector and the integration of digital technologies into almost every sphere of life exposes individuals and organizations to increasing cyber risks. Recently, excessive personal information has been breached, and substantial financial losses have occurred due to malicious attacks and other adverse events. For example, three billion personal records (all registered accounts) of Yahoo were breached in 2013 (Fung, 2017), and the credit card information of more than 100 million of Capital One's customers was hacked in 2019 (McLean, 2019).

Many corporations protect themselves against cyber risks, such as data breaches, by investing in security systems. Cyber insurance has been an alternative for managing risk and has grown rapidly compared to other insurance markets in recent years (Romanosky et al., 2019). However, although most experts agree that a significant part of future risks come from the digital world (PwC, 2019), the cyber-insurance market is still in its infancy and not as large as some industry commentators expected some years ago (Eling and Schnell, 2016). One reason for this might be that it is still unclear what drives cyber risk events.[1] How do firm-specific factors, such as size or industry, interact with the cost of cyber risk events?

The extant literature addresses this question by examining the interaction of firm-specific factors with the cost of cyber events (Romanosky, 2016; Eling and Wirfs, 2019; Aldasoro et al., 2020; Jung, 2021). However, all studies have focused on simple linear models that do not consider heterogeneity across the percentiles of the cost distribution.[2] This study aims to fill this gap using a quantile regression[3] that models the relationship between a set of

[1] In addition to this limitation, Biener et al. (2015) posit that asymmetric information and restrictive coverage limits may make it challenging to match supply and demand in the cyber insurance market.

[2] This might be problematic because research shows a large heterogeneity across different cyber risks (Eling and Wirfs, 2019). We hypothesize that the relations between financial costs of cyber events and firm-specific explanatory variables are not constant but vary across the quantiles of cost distribution. The study of Aldasoro et al. (2020) is the only one that mentions the impact of firm-specific factors on different quantiles of cyber costs; however, examining heterogeneity in cyber risks is not the focus of their study and is not discussed in detail.

[3] The quantile regression method allows fitting regression curves to the non-central locations of a response variable distribution, which the conventional linear regression method cannot achieve. This feature is significant to address the research question of this study in that 1) heterogeneous features of cyber risk drivers may exist depending on different costs, and 2) this heterogeneity may be associated

* Corresponding author.
E-mail addresses: martin.eling@unisg.ch (M. Eling), kwjung@postech.ac.kr (K. Jung), jeungbo.shim@ucdenver.edu (J. Shim).

https://doi.org/10.1016/j.insmatheco.2022.03.001
0167-6687/

firm variables and specific quantiles of the cost variable. The purpose is to identify the extent to which the relationship may vary across the quantiles of the cost distribution. In this context, to the best of our knowledge, we are the first to consider a dataset with individual security measures and their industry averages to examine the impact of firm-specific security levels on the costs of cyber events. All these results help identify potential risk exposures, develop an adequate pricing model for cyber insurance, and provide a comprehensive understanding of cyber risk (EIOPA, 2018).

Our dataset also allows us to examine whether the costs of data breaches can be approximated by the amount of breached data. The latter information is more readily available and has been considered by various researchers (Wheatley et al., 2016; Eling and Loperfido, 2017). However, the relationship between the number of breached records and the resulting financial loss has not been studied extensively and is still unclear.[4] The financial impact of a cyber loss event might vary according to the breach's extent, the entity's size (e.g., the number of employees or the amount of annual revenue), or the expected legal costs for a certain breach event.[5]

We use the results of the quantile regression models to calculate the cyber-insurance premiums for hypothetical corporate policyholders with heterogeneous characteristics. In this application, we compare the pure premiums of a quantile-based insurance pricing scheme with the premiums of a two-part generalized linear model (GLM) and the Tweedie model. The results indicate that the quantile model can help a cyber-insurer consider heterogeneity in the firm's characteristics, leading to more properly estimated premiums for potential insureds with different sizes, industries, and security levels.

Overall, we provide three main contributions. First, we use quantile regressions to understand better the heterogeneity of cyber losses across quantiles of the cost distribution. Second, we consider individual security measures and their impact on cyber costs. Third, we help better understand the relationship between the amount of data breached and the resulting financial losses.

The remainder of this paper proceeds as follows. Section 2 reviews the relevant literature related to cyber cost and loss modeling. Further, Section 3 describes the theoretical background on the methodology and descriptions of the variables and data used in the model. Moreover, Section 4 presents the empirical results of the quantile regression approach, followed by the descriptions of two other competitive models and applications to pure premium calculations in Section 5. Finally, Section 6 provides the conclusions and potential future research opportunities.

2. Literature review

To date, most cyber risk-related databases and studies have provided either data breach information—the number of breached records as a loss (Edwards et al., 2016; Eling and Loperfido, 2017)—or log file data (so called honeypots) that measure attack activities (Peng et al., 2017). However, neither of them incorporates the exact amount of financial costs caused by such risky events; hence, it has been challenging to examine the relationship between the size of a loss event (as measured by the attack indicators) and its financial impact.[6] Despite not having an adequate database, studies that empirically measure the financial impact of a cyber risk event have been presented in both academic and practical contexts. These studies implement either scenario-based/case study estimations (Lloyd's, 2017; Dreyer et al., 2018) or simple empirical regression-based models for the linear relationship between the financial cost of cyber loss events and firm-specific factors (Romanosky, 2016; Eling and Wirfs, 2019; Jung, 2021).[7]

Table 1 provides sectional comparisons between this study and seven previous empirical studies on estimating cyber costs. Unlike recent studies using the public dataset of the Privacy Rights Clearinghouse (PRC), in which only the number of breached records is available, the literature in Table 1 shows a selection of different data sources that allow us to examine the financial impacts of cyber risk events. In conducting their case studies, Lloyd's (2017) and Dreyer et al. (2018) focus on extreme cyber risk scenarios, in which the amounts of potential losses by a single cyber event are measured at the industry and national levels.[8] These studies find that cyber loss events have considerable potential to cause high economic costs for industry and society. They project a 250% increase in the loss ratio of the cyber-insurance industry for events in which extreme losses occur (Lloyd's, 2017, p. 48, for the case of a cloud service provider attack) and 1.27% of GDP by a systemic cyberattack (Dreyer et al., 2018, p. 28; considering The Netherlands as an example). However, their focus is primarily on the macro impacts of possible systemic cases, which might not be sufficiently useful for the insurance industry to understand the firm-specific drivers of cyber costs.

Eling and Wirfs (2019) extract cyber risk-related losses from an operational risk database provided by the SAS Institute and analyze them using an extreme value model, i.e., peaks-over-threshold (POT). First, they model the frequency and severity of losses with a Poisson distribution and POT, respectively. Subsequently, they use the dynamic EVT model to determine how economic losses are affected by risk and firm specifics.[9] The contribution of this study is in line with our goal—both explore the types of factors and

with different firm sizes. Our response variable (i.e., total costs) shows a long-tailed distribution on which one can expect cyber risk drivers may impact differently. Quantile regression can help capture this heterogeneity by estimating multiple rates of change (slopes) for all portions of a probability distribution of the response variable (Koenker and Bassett, 1978).

[4] Many cyber insurers have created algorithms to calculate the amounts of claims associated with loss events based on the extent of the breach, among others, but their construction and accuracy remain unclear. Academic literature has mainly used the Ponemon Institute (2019a, 2020) reports to translate the number of breached records to the costs in dollars (Jacobs, 2014; Edwards et al., 2016; Eling and Loperfido, 2017; Eling and Jung, 2018; Leong and Chen, 2020).

[5] For instance, Equifax—one of the largest credit bureaus in the United States—was hacked in 2017, and the personal information of 147 million people was breached. They were forced by The Federal Trade Commission (2020) to pay $425 million to assist people affected by this breach event.

[6] Lloyd's (2017) points out that traditional risk models in the insurance industry have been based mostly on authoritative sources of information—government agencies and industry—but no case has been made for modeling cyber risk, which relies primarily on Internet sources. Nevertheless, many public and private databases have incorporated significant numbers of observations collected from announcements of governmental agencies and verifiable media sources (e.g., the Privacy Rights Clearinghouse (PRC), Advisen). The dataset used in this study was obtained from Cowbell Cyber, Inc., a United States-based cyber insurer with its risk analytics; it collects data from the public (e.g., PRC) and private (e.g., Advisen) sources.

[7] Scenario-based studies or case studies examine how the consequences of a proposed model appear when hypothetical profiles of cyberattacks (e.g., in a certain country (The Netherlands) or a certain extreme risk event, such as a blackout) are applied. However, the literature with simple empirical, regression-based models investigates empirical datasets with information on loss observations (i.e., the actual financial cost of cyber events and firm-specific factors).

[8] Lloyd's (2017) assumes two hypothetical scenarios: 1) an attack on cloud service providers; 2) an attack on an operating system used extensively in the global market. The report analyzes how the consequence of the loss amounts and changes in loss ratio would appear in both large and extreme loss cases when attacks occur. Dreyer et al. (2018) propose a model to measure the cost of cyber risk events, and they assume high-profile cyberattack cases to derive the potential cost of loss events.

[9] Eling and Wirfs (2019) consider as risk-specific systems and technical failures the failed internal processes and external events in which differentiation between malicious and negligent types is not fully clear. They are considered firm-specific regions, industry, and the number of employees representing the firm size. For the industry-specific factor, they categorize observations into the financial services and non-financial industries, and our approach is in line with their work.

Table 1
Comparison of Relevant Literature with This Study in Cyber Cost Estimation.

	Romanosky (2016)	Lloyd's (2017)	Dreyer et al. (2018)	Eling and Wirfs (2019)	Aldasoro et al. (2020)	Palsson et al. (2020)	Jung (2021)	This study
Data	Advisen (2005-2014)	Cyence	Own data	SAS OpRisk	Advisen	Advisen	Cowbell Cyber	Cowbell Cyber
			(2010–2017)	(1995–2014)	(2002–2018)	(2000–2018)	(2005–2018)	(1992–2020)
Observations	265	-	590	1,579	3,228	75,000	11,412	21,555
Method	Linear regression	Scenario-analysis	Case study	Dynamic extreme value theory	Linear, quantile, and probit regressions	Random forests	Linear regression	Quantile regression
Focus of study	A simple approach with an average impact	Exploration of hypothetical extreme scenarios	Sectional (industry) exposures to cyber loss events	Cyber loss modeling and firm-specific characteristics for extremes	Cyber cost modeling with firm- and industry-specific characteristics	Analysis of cyber loss events and their financial impact	Extreme data breach loss modeling and cost effect	Exploration of quantile effects in cyber cost estimation
Main findings	Breach size is positively associated with economic cost	Cyber loss events can lead to a 250% increase in industry loss ratio	A systemic cyber event in the Netherlands can lead loss size to 1.25% of GDP	A larger financial firm is more exposed to extreme losses than a smaller non-financial firm	Average cost of cyber events increases and is higher for larger firms	Malicious risk type dominates most incidents, and the financial industry is most exposed to cyber risk	The smaller the loss size, the higher cost per record (in accordance with median)	Heterogeneity in the relationship between cyber cost and breach size

Note: SAS is a US-based software firm also collecting data on operational risk; Advisen is a US-based data provider collecting and integrating cyber incident data; Cyence is a US-based cyber risk modeler part of Guidewire Software; Cowbell Cyber is a US-based cyber insurer integrating databases from public and private sources with their cyber risk analytics.

how they affect the size of economic costs by cyber risk events at extreme quantiles. However, they focus mainly on cyber risk events, whereas our dataset enables the assessment of the impact of the size of data breach events, which is of primary interest to cyber insurers. This is because data breach losses are one of the main types of losses covered in stand-alone cyber-insurance policies (Romanosky et al., 2019). Additionally, the dynamic EVT model employed in this study leads to the primary examination of large events on the right tail. Our approach helps examine the entire spectrum of the cost distribution and thereby determine the heterogeneous impacts of loss drivers.

Romanosky (2016) investigates the impact of cyber risk events on their actual costs using the Advisen dataset[10] by simple regression approach, which incorporates firm-specific factors to address the relationship between the number of breached records and financial costs. The author demonstrates that the number of breached records is highly correlated with the cost of breach events—a 10% increase in the number of breached records corresponds to a 2.9% increase in cost, indicating that larger breaches result in greater costs. By splitting observations with their median value, Jung (2021) finds that a smaller breach event can generate a higher cost per record. However, both studies are limited—they reveal an average impact of cyber risk events known to be heavy-tailed (see Maillart and Sornette, 2010; Wheatley et al., 2016; Eling and Jung, 2018).

Recently, Palsson et al. (2020) used the Advisen database to conduct 1) a qualitative analysis of cyber loss events by industry and the firm's size and 2) a quantitative analysis of the financial cost of cyber incidents by the type of risk. Although they determine the factors affecting the financial cost of breached entities, their approach is at the aggregate level to classify observations into three different cost intervals, revealing which factor has the highest impact. It does not provide any details on how the size of the breach and firm-specific factors can drive the financial costs of events, which the authors recognize as a limitation.

Aldasoro et al. (2020) also use the Advisen database to determine the factors that can influence the cost of cyber loss events. They find that larger firms have higher cyber costs and that incidents affecting multiple entities result in additional costs. They also investigate the effects of cloud services and crypto-related activities on cyber costs. They mention an analysis of the impact of firm-specific factors on different quantiles of cyber costs; however, it is not the primary focus of their study. It was not discussed in detail.

For the insurance industry, regarding the pricing of cyber insurance and estimation of claims related to cyber risk events, it is important to determine whether heterogeneous quantile effects of firm-specific factors and breach sizes exist on the size of cyber economic losses and, if so, how they appear. In this sense, quantile effects first account for the average effects of cost events examined by most studies. The effects also address the heterogeneity that can help understand better how different the impacts of loss drivers on the cost of cyber events are and, subsequently, how this heterogeneity can be associated with insurance premiums. The literature reviewed in this section has not examined this aspect. None of the reviewed studies considered security information at the firm level because such information is not available in datasets.

[10] The database provided by Advisen recently has been used in the cyber risk context to examine the features of cyber loss events (see Aldasoro et al., 2020; Palsson et al., 2020). The database offers a large set of observations on cyber loss events and their information, for example, firm characteristics (i.e., the sizes of companies and the types and numbers of employees) and loss-specific information (i.e., source, type, and amount). The total size of our dataset is smaller than Advisen's dataset, but it includes additional security-metric variables and various data sources, as shown in Section 4.1.

3. Methodology, variables, and data

3.1. Methodological background of quantile regression

A simple approach to predicting the relationship between the variables of interest is a regression model that features linearity with the Gaussian distributional properties of observational noise. The least squares estimation of the linear regression model with a Gaussian error distribution is tractable, particularly for supporting optimal unbiased estimators. The approach focuses on conditional mean surfaces, and it is easy to interpret the effects of the variables. However, the effects estimated with this conventional approach do not indicate the exact genuine effects of outliers or observations away from the mean (Koenker and Hallock, 2001). In particular, there is a need for a case with a long-tailed distribution to be addressed by robust alternatives to the conventional approach, which is limited in cases with non-Gaussian errors.

To address this limitation, Koenker and Bassett (1978) propose a quantile-based regression, which helps understand the relationship between the response variable and the key variable of interest, for example, in the tails. This implies that the effects of covariates on observations over the median or mean can be significantly different from those at extreme quantiles. This method estimates and inference about conditional quantile functions, not conditional mean functions; hence, it can facilitate the prediction of the linear relationships among variables at any quantile.

The general form of the quantile regression model is presented here. If the linear regression model is set as $y_i = x_i'\beta + u_i$ for $i = 1, \ldots, n$, where y_i and x_i are redefined hereafter as the response variable and the k-dimension vector of covariates, respectively, then the least squares estimator specifies the conditional mean function of the response by minimizing the sum of the squared residuals. That is, $\sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$, where β is a vector of the parameters to be estimated. However, this conditional mean estimation method does not allow us to investigate the possible heterogeneity of the parameters across the conditional distribution of the response variable. The quantile regression is appropriate when, rather than being constant, the effects of the covariates on the response vary at different points of the conditional distribution of the response variable. For a different quantile—$\theta \in [0, 1]$—the basic quantile regression model can be written as follows:

$$\begin{aligned} &y_i = x_i'\beta + u_i, \quad i = 1, \ldots, n \\ &Q_\theta\left(y_i \mid x_i\right) = x_i'\beta_\theta, \end{aligned} \tag{1}$$

where $Q_\theta\left(y_i \mid x_i\right)$ denotes the θ^{th} conditional quantile of y_i on the covariates x_i. Parameter β_θ varies with different quantiles θ. The parameter β_θ can be estimated by solving the conditional quantile objective function:

$$\hat{\beta}_\theta = \operatorname{argmin}_{\beta \in R} E\left[\rho_\theta\left(y_i - x_i'\beta_\theta\right)\right], \tag{2}$$

where $\rho_\theta(\cdot)$, which is known as the check function, is defined as follows:

$$\rho_\theta\left(u_{\theta i}\right) = \begin{cases} \theta u_{\theta i} & \text{if } u_{\theta i} \geq 0 \\ (\theta - 1)\, u_{\theta i} & \text{if } u_{\theta i} < 0 \end{cases} \tag{3}$$

Similar to the least squares estimator, conditional quantile estimators are produced by minimizing the sum of the weighted absolute residuals. Thus, estimator $\hat{\beta}_\theta$ is a solution of

$$\min_{\beta \in R} \left[\sum_{i: y_i \geq x_i'\beta_\theta} \theta \left| y_i - x_i'\beta_\theta \right| + \sum_{i: y_i < x_i'\beta_\theta} (1 - \theta) \left| y_i - x_i'\beta_\theta \right| \right] \tag{4}$$

Equation (4) is solved by linear programming methods, and the bootstrap method is used to estimate the standard errors of the parameters. The Markov chain marginal bootstrap (MCMB) is implemented to construct confidence intervals for the regression quantile estimates. For quantile regression, the MCMB method has an advantage—it solves p one-dimensional equations instead of p-dimensional equations.[11] Unlike the ordinary least squares (OLS) regression, the quantile regression approach is robust to extreme observations in the response variable but not to extreme points in the covariate space (Koenker and Hallock, 2001).

3.2. Description of the variables

We consider a new large dataset that provides several firm-specific factors. The following elaborates on each variable used in this study. Table 2 summarizes the definitions of the variables.

Total costs

The total cost variable is the target in our model, which accounts for how much is finally paid to cover the loss due to a cyber risk event. This variable incorporates the financial losses that result from cyber incidents. Some observations have values identical to those of the financial damage variable. The observations different from financial losses include litigation costs and other costs, apart from any incidents that cause direct damage. Litigation costs for a loss event resulting from cyber risks comprise, for example, costs from a class action lawsuit, judicial rulings, retainment of a legal counsel, and fines imposed by government agencies (Romanosky, 2016). Including additional costs related to cyber risk, events are important to explain why this variable should be the target because most current cyber-insurance policies cover liabilities, data compromise responses, and defense costs (Romanosky et al., 2019). We take the natural logarithm for the variable to avoid highly skewed observations; hence, the log-transformed variable can appear to be normally distributed.

Financial damage

This variable represents the size of the pure financial damage caused by a cyber risk event. It does not include any other costs, such as the cost of litigation. Thus, it cannot represent the exact amount of the economic cost that a breached entity should bear, which leads us not to consider this issue in the main approach. Nevertheless, this variable can help check whether the impacts of factors on the amount of the loss are robust and, if not, how they differ. We take the natural logarithm of the total cost variable.

Records

The record variable shows the number of breached records per event, and it is a key explanatory variable in our model (particularly for the claims cost calculation in Tables 7 and 8). The breached records are the personal information that organizations manage in their databases. The core rationale for the relationship between the number of breached records and the total costs is that the number of breached records is the key determinant of cyber-insurance coverage in the market. Values of breached records can vary depending on firm-specific and industry factors. The expected impact of this variable on the total costs is positive; thus, the larger the size of the breach, the larger the total costs. In addition

[11] In terms of statistical inference on the model parameters with a small sample size and for more extreme quantiles, literature (e.g., Kocherginsky et al., 2005; Tarr, 2012) has documented the efficacy of the bootstrap method for quantile regressions. Specifically, Kocherginsky et al. (2005) find that the bootstrap approach is reliable for inferring the quantile regression method with a small sample size (200, 400, and 500 sample sizes tested) and small size bootstrap replications. Tarr (2012) also shows the effectiveness of several bootstrap methods based on the study of Kocherginsky et al. (2005) for the sample size less than 200 to construct confidence intervals.

Table 2
Definition of the Variables.

Variable	Definition	Type
Total costs	Total financial loss caused by a cyber risk event, including the cost itself by the event and the litigation cost if applicable	Numeric
Financial damage	Size of pure financial damage by a cyber risk event that does not consider any other cost (e.g., the litigation cost)	Numeric
Records	Number of personal information records breached or leaked by a risk event (number of breached records)	Numeric
Revenue	Annual revenue size of a breached entity	Numeric
Number of employees	Total number of employees working for a breached entity	Numeric
Malicious	Indicator = 1 for observation with malicious intention event and 0 otherwise	Binary
Negligent	Indicator = 1 for observation with negligent risk event and 0 otherwise	Binary
Network/Processing error	Indicator = 1 for observation with network disruption or processing error event and 0 otherwise	Binary
Financial	Indicator = 1 for observation in the financial service industry and 0 otherwise	Binary
Health	Indicator = 1 for observation in the health care/medical industry and 0 otherwise	Binary
Wholesale/Retail	Indicator = 1 for observation in the wholesale/retail industry and 0 otherwise	Binary
Manufacturing	Indicator = 1 for observation in the manufacturing industry and 0 otherwise	Binary
Information	Indicator = 1 for observation in the information industry (e.g., media, software publishers, data services) and 0 otherwise	Binary
Administrative/supporting management	Indicator = 1 for observation in the wholesale/retail industry (e.g., credit bureaus, employment services) and 0 otherwise	Binary
Yr2014	Indicator = 1 for observation occurring after 2014 and 0 otherwise	Binary
Relative security	Indicator = 1 for observation with a lower individual security score (scale 0–100) than the average security score of firms in the same industry and 0 otherwise	Binary

Note: Numeric variables are log transformed in the model.

to investigating this expected positive relationship, we examine how this relationship can vary at different quantiles of the cost distribution.

Jung (2021) differentiates the effects of the size of the breach in the loss amount by modeling the dataset separately above and below the median value. The author determines that the cost per record of breaches is greater for breach events below the median value, whereas each breached record has less impact on the amount of loss for events above the median value. However, it cannot provide evidence on how these impacts are heterogeneous across quantiles, with extant findings that a cyber loss distribution is heavy-tailed and has a wide spectrum. Our design and approach can offer a more reasonable answer to this question. We use this variable as the key to explaining how the claims can be calculated with other firm-specific factors when a breach event occurs. We also take the logarithmic transformation for this variable to alleviate highly skewed observations.

Revenue

The annual revenue of an organization is also a key variable that can be used to explain how large its exposure is to a cyber risk event. Romanosky et al. (2019) analyzed 100 policies traded in the cyber-insurance market. They found that applications for insurance contracts begin by collecting information about the firm (industry, business operations), including its financial status and revenue. Revenue can represent the firm's size and be associated with risks exposure as high-profile firms can be the targets of cyberattacks.[12] Again, we transform this variable logarithmically to alleviate highly skewed observations.

Number of employees

The number of employees indicates how large an organization is and, thus, how much it is exposed. This variable can be argued that revenue and the number of employees can positively correlate because both can indicate the firm's size. This scenario is partially correct, particularly for a traditional secondary industry where manufacturing and production are the core of the business. However, cyber risk events are more likely to be experienced by unicorn tech firms[13] or businesses that operate exclusively online (Jung, 2021), where the traditional business structure may not be fully appropriate.[14] Rather, we posit that the number of employees can be used to determine the degree of exposure. This is because the larger the number of employees in the corporate network, the more endpoint nodes are exposed to cyberattacks. To examine this argument, we use this variable to check robustness by replacing the revenue size representing the firm's size and its risk of cyber exposure.[15] This variable is also transformed logarithmically in the model.

Type of risk

Whether a loss occurred due to malicious action is another important factor to assess the potential cost. This is because a malicious action is more likely to affect multiple entities/regions in the interconnected network, whereas negligent action is likely

[12] Several cyber insurers have diversified their offers with differentiated insurance prices by revenue to indicate the firm's size. For example, Cowbell Cyber provides coverages for small, middle-sized, and large firms separately by assessing heterogeneous security levels, revenue, and degree of exposure. Romanosky et al. (2019) also claim that current policies in the market are priced with the greatest weight on the base asset value or revenue.

[13] *Unicorn* was first used by Aileen Lee, a venture capitalist and the founder of Cowboy Ventures (Lee, 2013). A *unicorn firm* is defined as a private start-up with a value exceeding 1 billion. This type of firm is regarded as a successful player in the venture capital industry; hence they appear to have already exited the venture capital market (Lee, 2013). Many tech-oriented start-ups with online-based business concentration have joined the Unicorn Club referred by Aileen Lee, the firms that are, for example, Facebook and LinkedIn.

[14] In the PRC dataset with the 15 most extreme cases over the last decade, Jung (2021) identifies that 66.7% of the cases occurred in online-based businesses (62.5% of the total breached records).

[15] With the variance inflation factor having a value of 4.2496, with 5 being the threshold that determines the presence of multicollinearity, we find no evidence of multicollinearity between the amount of revenue and the number of employees. Therefore, we do not use them in one model; instead, we apply them separately to check the robustness of the model.

to be limited to a single loss event.[16] In addition to this aspect, which has already been considered in the literature, our dataset allows us to examine the potential impact of other types of risk—network disruption and processing errors. This risk type differs from others—it aims primarily to interrupt business processes in an interconnected network environment. To summarize, we construct three binary variables to test the impacts of various types of risk, where malicious data breaches, negligent data breaches, and network disruption/processing errors are distinguished. Variables of risk type assign 1 to the corresponding risk-type losses and 0 otherwise. The dataset in this study includes the ten types of risks listed in Table 2, where one type (unintended data breach) belongs to a negligent risk, and two types (network disruption and configuration/processing errors) can be classified as network/process errors.

Industry

Financial service industries are more exposed to cyber risk events in terms of frequency and severity. The Ponemon Institute (2019a) finds that the financial industry has one of the highest average total costs and costs per record of data breaches. This finding has been reported since 2005.[17] Therefore, it is important to examine whether risk events to entities in the financial services industry are more likely to have a larger impact on the size of costs. The result might interest insurers in considering how industry determinants can affect the premium calculation. This variable takes a value of 1 for financial entities, including banks, insurers, and real estate rental/leasing agencies and 0 otherwise.

Like the financial services industry, one of the most targeted industries recognized in the study is the information industry, administrative/supporting management entities, and the health care/medical industry (Dreyer et al., 2018; Ponemon Institute, 2019a). These industries are at risk for personal credit information, data storage, and biometric information.[18] Additionally, we consider the wholesale/retail and manufacturing industries because a relatively sufficient number of entities belong to this dataset. Each variable in an industry is assigned a value of 1 for observations associated with the corresponding industry and 0 otherwise.

The year 2014

This variable shows a possible time break in 2014 in the dataset by assigning 1 to observations after 2014 and 0 otherwise. It can divide the data period structurally based on the heterogeneous characteristics of the loss trend. Jung (2021) claims that there was a potential time break in the cyber loss trend in 2014, and three statistical tests support this claim.[19] The author uses the dataset (Cowbell Cyber) of this study and a public dataset from the PRC, showing nearly identical time trends in Fig. 1. As we also use the same dataset for the model in this study, there is a need to explore how influential a potential time break in 2014 can be in the amount of cyber costs.

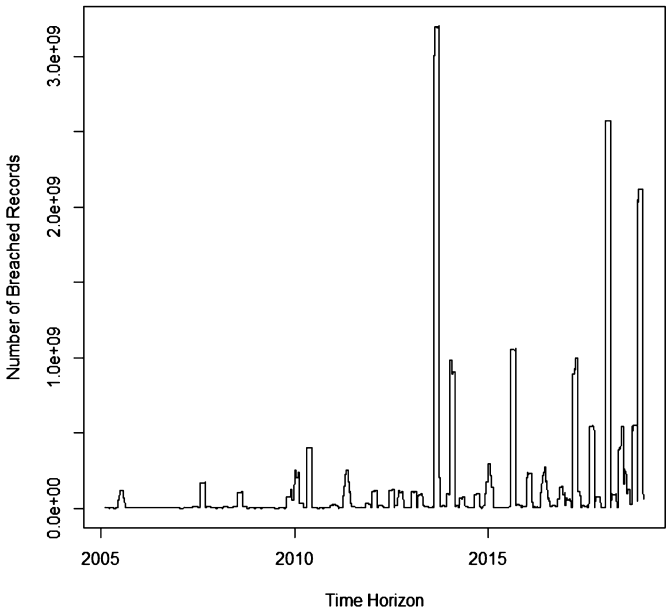

Fig. 1. Time series plot of loss severity (the number of breached records) between 2005 and 2018. The trend shows an increased number of losses of this dataset since 2014, which has already turned out to be statistically significant in the study by Jung (2021).

Security level (*relative risk exposure*)

There are unique security measures in our dataset that have been developed by the data provider (Cowbell Cyber). The measures are the outcome of metrics that quantify the overall security level of an organization[20] and its industry average security level (i.e., the average of individual security measures in the same industry; see Table 4 for more details). These measures can be interpreted so that the higher its value, the better a cyber risk posture for the corresponding organization or industry (i.e., the better its cyber security system, the lower the overall cyber risk exposure). To the best of our knowledge, this type of information representing the level of risk exposure or cyber security has never been considered in modeling the economic costs of cyber risk events. This information can help understand how a firm's security level and risk exposure interact with the cost resulting from cyber risks.

To properly embed this information in the model, we use these measures to create a variable (relative security in Table 2) representing whether a victim can be a weaker security point than the others in the industry. An indicator variable equal to 1 for firms with lower individual security measures than the industry average

[16] Evidence can be found in the literature (see Edwards et al., 2016; Eling and Loperfido, 2017; Jung, 2021) where statistics and estimates of malicious loss events, generally, are larger than the losses associated with negligent events.

[17] Supporting this finding of Ponemon Institute (2019a), Eling and Wirfs (2019) also posit that the financial industry has higher exposure to cyber risks than other industries. Hence, they also included a binary variable for inclusion in either the financial or non-financial industry.

[18] This industry is important because it has many observations, including notable loss events. For instance, Target, a retail giant in the US, breached nearly 41 million of its customers' payment card accounts, and the company paid approximately $18.5 million for a multistate settlement (McCoy, 2017). Another extreme case is Marriott International, which breached the records of more than 300 million customers, including credit card and passport IDs, and the company was fined £18.4 million by the UK Information Commissioner's Office (ICO) for violating the EU General Data Protection Regulation (Tidy, 2020).

[19] Jung (2021) uses additional information to provide two supporting grounds on the identified break. One is a seemingly accelerated progress in information technology represented by Moore's law that the number of transistors in an integrated circuit doubles every two years. The author finds a structural change in technological development in 2014 using the data of the number of transistors on microchips, and the shift leads to doubling the number every 18 months. Additionally, the study identifies that ten of the 15 most extreme data breach loss events over the last decade have occurred in online-based business entities. The other rationale is the enforcement of the laws concerning the notification of data breaches in several states in the US over the last decade, possibly driving the increase in cases involving reported losses across the nation.

[20] The individual measure is the average of the following factors scaled between 0 and 100—network security, cloud security, endpoint security, dark intelligence, transfers of funds, cyber extortion, compliance, and insider threats. The factors are described as follows. Network security measures the strength of the organization's network infrastructure; cloud security indicates an organization's cloud security risk and use of public cloud/storage; endpoint security stands for endpoint vulnerabilities (servers, mobile, Internet of Things) and security measures implemented; dark intelligence shows an organization's exposure to the darkweb; funds transfer is risk markers related to spear phishing (email, SMS, or web); cyber extortion indicates potential exposure to extortion-related attacks (ransomware); compliance stands for the level of compliance to security standards.

Table 3
Types of Cyber Risks.

Malicious risk	*Negligent risk*	*Network/process error risk*
• Cyber extortion (92) • Identity theft (382) • Malicious data breach (5,576) • Phishing and spoofing (649) • Privacy breach (10,905) • Skimming (188) • Stolen data (1,128)	• Unintended data breach (1,600)	• Configuration/Processing errors (440) • Network disruption (519)

Note: The figures in parentheses account for the sample size of each risk type. The sum of the sample sizes in this table is the size of the full data set minus the number of unknown types, which is not considered in dummies of risk types.

Table 4
Industry Average Measures for Security/Risk Exposure.

NAICS code	23 (Construction)	31 (Manufacturing)	32 (Manufacturing)	33 (Manufacturing)
Value	66	59	64	56
NAICS code	42 (Wholesale Trade)	44 (Retail Trade)	45 (Retail Trade)	48 (Transportation)
Value	61	57	44	57
NAICS code	51 (Information)	52 (Finance and Insurance)	53 (Real Estate and Leasing)	54 (Scientific, Technical Services)
Value	44	43	63	58
NAICS code	55 (Management of Companies)	56 (Admin, Support Management)	61 (Educational Services)	62 (Health Care and Social Assistance)
Value	63	49	52	60
NAICS code	71 (Arts and Entertainment)	72 (Accommodation and Food Services)	81 (Other Services)	92 (Public Administration)
Value	65	58	63	55

Note: The value of cyber risk posture in this table is interpreted so that the higher the value, the better the average cyber risk posture of an industry. The value is measured based on the following factors: network security (the strength of the organization's network infrastructure); cloud security (an organization's cloud security risk and use of public cloud/storage); endpoint security (endpoint (servers, mobile, Internet of Things) vulnerabilities and security measures implemented); dark intelligence (an organization's exposure to darkweb); funds transfer (risk markers related to spear-phishing via email, SMS, or web); cyber extortion (potential exposure to extortion-related attacks, e.g., ransomware); compliance (level of compliance to security standards).

Table 5
Descriptive Statistics.

	N	Mean	St. Dev	Max	Q3	Median	Q1	min
Total costs	3,162	2.941m	40.930m	2.000b	8,000	0.000	0.000	0.000
Financial damage	745	9.123m	80.526m	2.000b	1.750m	195,000	13,247	0.000
Records	12,280	1.436m	41.817m	3.000b	1,000	28.000	1.000	1.000
Revenue	21,317	19.575b	51.153b	514.00b	11.539b	192.67m	26.963m	0.000
Num. employees	21,409	47,899	204,588	2.769m	22,899	1,300	187.00	0.000
Malicious	21,555	0.848	0.359	1.000	1.000	1.000	1.000	0.000
Negligent	21,555	0.074	0.262	1.000	0.000	0.000	0.000	0.000
Processing error	21,555	0.044	0.206	1.000	0.000	0.000	0.000	0.000
Financial	21,555	0.283	0.451	1.000	1.000	0.000	0.000	0.000
Health	21,555	0.059	0.235	1.000	0.000	0.000	0.000	0.000
Retail	21,555	0.093	0.290	1.000	0.000	0.000	0.000	0.000
Manufac	21,555	0.051	0.220	1.000	0.000	0.000	0.000	0.000
Information	21,555	0.114	0.318	1.000	0.000	0.000	0.000	0.000
Admin	21,555	0.218	0.413	1.000	0.000	0.000	0.000	0.000
Yr2014	21,289	0.596	0.491	1.000	1.000	1.000	0.000	0.000
Relative security	21,511	0.406	0.491	1.000	1.000	0.000	0.000	0.000

Note: Total costs and revenue are in the dollar unit. Total costs and financial damage include several blanks excluded in this table. In particular, the variable of total costs contains more zeros than financial damage, leading to larger statistics for financial damage, although total costs are always higher or equal to financial damage when comparing observation-by-observation. As noted in Table 3, the "unknown" risk type is not considered in variables of risk types because no sufficient information to classify observations of this type is offered in the dataset; "b" and "m" stand for billion and million, respectively.

and 0 otherwise is included to control cost differences between firms with different security levels. An individual security measure lower than the industry average implies the negative externality of the cyber security landscape. In the landscape, cyber attackers may change their targets to a relatively weaker point in the same industry with high interconnection, thereby causing an increase in the probability of a loss and the amount of the cost for victims (Zhao, Xue and Whinston, 2013). The relationship between the total costs and the relative security level in our model can help evaluate the association of cyber security externality with the size of cyber losses.

3.3. Data

The data are obtained from a cyber-insurer—Cowbell Cyber Inc.—which has collected 21,555 observations from public data sources (e.g., from PRC) and private datasets (e.g., Advisen, DarkOwl, MS Secure Score, and FICO) across the US between 1992 and 2019. It offers information on the amount of revenue, the number of employees, North American Industry Classification System (NAICS) codes, state, event date (exact date of an incident), type of risk, amount of financial damage, total costs, and the number of breached records. The total costs can be identical to or higher than the amount of financial damage, depending on whether there is litigation. Our model considers the total costs to address both the first- and third-party losses of breach events.

Among the entire dataset, we find 933 observations with nonzero total costs and 324 cases with the amount of financial damage and the corresponding number of breached records for the data period. We use the dataset with 933 observations to examine the impacts of firm-specific factors on the total costs without the breach size (i.e., ex-ante loss analysis for insurance pricing). However, the dataset with 324 observations is used to investigate the relationship between the size of the breach and the amount of the loss in addition to firm-specific effects (i.e., ex-post loss analysis for the claims cost calculation).[21] The raw data categorize cyber risks into 12 types; however, the filtered dataset contains nine types of risks, where most risks are based on malicious intentions. We categorize nine risks into three broad types: malicious, negligent, and network disruption/processing errors (see Table 3).[22] To account for the types of industry, we add six binary variables (see Table 2 for definitions of the variable).

Table 5 shows descriptive statistics of our full dataset with additional variables. It is observed that most of our dataset occurred by malicious action. Concerning industry variables, approximately 30% of the events occurred in the financial service industry, in which the banking, insurance, real estate, rental, and leasing industries are included. The administrative/supporting management industry has the second largest number of observations (approximately 21%). Further, approximately 60% of observations occurred after 2014, during which the expectation was that more extreme cases would be observed (Jung, 2021). Finally, approximately 40% of the observations have weaker security levels than the average security level of each industry.

[21] The small sample size of actual loss observations is still a considerably typical limitation for cyber risk studies, especially because victims are reluctant to disclose details of their losses due to reputational concerns (Eling and Wirfs, 2019; Aldasoro et al., 2020). We note that the quantile regression method has a relatively small sample size in many research fields. For example, a small sample size issue is a typical concern in healthcare research. Revzin et al. (2014) address this with a quantile regression method to analyze six different experimental studies, including 35 to 60 sample sizes.

[22] Malicious and negligent types of risk are differentiated by whether an actor's (malicious) intention is inherent in a risk event. The network disruption/processing error type incorporates internal process failures or system operation failures, which may or may not be caused by intentional/unintentional actions.

4. Empirical results

We implement quantile regressions for every 5^{th} quantile starting from the 5^{th} and ending with the 95^{th}. However, we only present the results of five distinctive quantiles (5%, 25%, 50%, 75%, and 95%) in Tables 6 and 7 to conserve space. Lower quantiles (5% and 25%) of the total cost distribution represent smaller cost events, whereas higher quantiles (75% and 95%) represent higher cost events. Table 6 excludes the number of breached records, which is not known as ex-ante but can be used as ex-post to calculate the cost of claims. Table 7 includes the number of breached records as explanatory variables to reflect this aspect. Consequently, Table 7 provides information about the potential estimation of claims amounts when a breach occurs. In Tables 6 and 7, we include firm revenue in the regression model to represent the firm's size (the degree of exposure to cyber risk).

4.1. Cyber-insurance pricing

Table 6 shows that the coefficients of a firm's revenue are not statistically significant at the 5% quantile level, indicating that extremely low-cost events are not associated with the revenue of an organization.[23] However, we observe that there is a significant positive relationship between a firm's revenue and the total costs across the remaining quantile levels of the cost distribution. Notably, the magnitude of the coefficients becomes larger (smaller) in the lower (upper) quantiles, suggesting that the impact of a firm's revenue is stronger (weaker) in the lower (upper) levels of total costs.

Our raw data demonstrate that the amount of revenue and number of employees of firms above the 90% quantile of total economic costs are, on average, 2.21 and 2.78 times higher than those below the 10% quantile of total economic costs, respectively. In other words, firms in the lower economic cost quantiles are more likely to be small or medium-sized enterprises (SMEs), whereas those in the higher economic cost quantiles are relatively large. Therefore, our finding implies that small firms can have a higher cost per risk exposure. Mispricing issues may arise if firms in the lower or upper levels of cost distribution are priced based on the average effect on the central location of the cost distribution.[24]

Although we provide no explicit evidence on the rationale behind this finding of the smaller impact of firm size in the upper quantiles, one possible explanation is the role of economies of scale. Small firms tend to lack the capability to make sufficient investments in their cybersecurity systems. They face challenges in enhancing cyber security due to limited budgets and difficulties in hiring skilled personnel (Aiyer et al., 2021). Moreover, many IT-related costs (e.g., the cost of hardware, software, and key personnel) are fixed rather than variable, leading larger firms to take advantage of economies of scale (Gordon et al., 2018).

As Gordon et al. (2018) argue, many firms need to comply with reporting requirements, such as the requirement to file reliable financial reports with the United States Securities and Exchange Commission (SEC). Compliance with these regulations also requires cybersecurity investments and documentation. The Sarbanes-Oxley Act (SOX) of 2002 and the 2011 SEC Disclosure Guidance requiring firms to disclose their cybersecurity risks and incidents put heavy weight on providing reliable financial reports to the public. Gordon and Smith (2007) and Gordon et al. (2015) claim that

[23] This finding can provide the usefulness of the quantile-based method by comparing it with the result of Romanosky (2016), which indicates that the firm size represented by the annual revenue is positive and significant in explaining the average cost of cyber risk events.

[24] The premium estimation with the quantile regression is discussed in Section 5, which indicates how to determine the quantile for pricing.

Table 6
Quantile Regression Results for Insurance Pricing (without the Number of Breached Records).

	Response variable = ln(total costs)					
	QT=0.05	QT=0.25	QT=0.50	QT=0.75	QT=0.95	OLS
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
ln(revenue)	0.0484	0.2355***	0.2199***	0.2182***	0.1569***	0.1596***
	(0.0471)	(0.0394)	(0.0399)	(0.0363)	(0.0297)	(0.0338)
Malicious	-2.4448***	-1.6494***	-1.1879**	-0.5673	-0.4982	-1.3426***
	(0.7134)	(0.5967)	(0.6044)	(0.5502)	(0.4504)	(0.5121)
Negligent	-0.8591	0.3448	-0.8058	-0.9565	-0.5615	-0.5041
	(1.0454)	(0.8743)	(0.8856)	(0.8062)	(0.6600)	(0.7504)
Processing	-1.7892**	-0.6085	-0.6860	0.0875	1.0792**	-0.6330
	(0.8421)	(0.7043)	(0.7134)	(0.6494)	(0.5316)	(0.6045)
Financial	-0.9986**	-1.7507***	-1.1556***	-0.1209	0.4033	-0.7248**
	(0.4369)	(0.3654)	(0.3701)	(0.3369)	(0.2758)	(0.3136)
Health	1.6141**	0.2463	-0.7212	-0.8123	-1.2631**	-0.2364
	(0.7807)	(0.6530)	(0.6614)	(0.6021)	(0.4929)	(0.5604)
Retail	-0.9816*	-0.7058	0.0686	0.0540	0.0437	-0.1693
	(0.5390)	(0.4508)	(0.4566)	(0.4157)	(0.3403)	(0.3869)
Manufac	-0.4995	-1.1746**	-0.6394	-0.1974	0.6645	-0.4406
	(0.6667)	(0.5576)	(0.5648)	(0.5142)	(0.4209)	(0.4786)
Information	-0.6439	-0.9978**	-1.3617***	0.2316	0.0702	-0.5612*
	(0.4745)	(0.3969)	(0.4020)	(0.3660)	(0.2996)	(0.3406)
Admin	-0.6725	-1.9022***	-2.0310***	-0.1961	0.2673	-1.0104**
	(0.5731)	(0.4793)	(0.4855)	(0.4420)	(0.3618)	(0.4114)
Yr2014	0.6822**	0.9018***	1.3224***	0.9878***	0.6309***	0.9524***
	(0.3255)	(0.2723)	(0.2758)	(0.2510)	(0.2055)	(0.2337)
Relative security	0.5746*	0.1551	0.4097	0.3179	0.2297	0.4699*
	(0.3346)	(0.2799)	(0.2835)	(0.2581)	(0.2113)	(0.2402)
Constant	8.5246**	7.5272***	9.1358***	10.3803***	13.5443***	10.1854***
	(1.0950)	(0.9159)	(0.9277)	(0.8445)	(0.6913)	(0.7861)

Note: This table shows the results of our baseline quantile regression on the response variable, the total costs by cyber risk events. We present estimates and standard errors (in parentheses) of five quantiles (QT) and the OLS model. The independent variables considered in the model are the revenue, three risk types, affiliation to six different industry sectors, a 2014 time break, and the firm's relative security level. *, **, and *** indicate that the p-value is less than the significance levels, 10%, 5%, and 1%, respectively, and coefficients without any indicator show no statistical significance. We use SAS to implement the quantile regression models; the number of observations is 933 after excluding missing values.

Table 7
Quantile Regression Results for Claims Cost Calculation (with the Number of Breached Records).

	Response variable = ln(total costs)					
	QT=0.05	QT=0.25	QT=0.50	QT=0.75	QT=0.95	OLS
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
ln(records)	0.2291***	0.3431***	0.2601***	0.2114***	0.1500***	0.2518***
	(0.0258)	(0.0448)	(0.0357)	(0.0267)	(0.0339)	(0.0308)
ln(revenue)	-0.0372	0.0264	0.0574	0.1073***	0.0905*	0.0471*
	(0.0366)	(0.0635)	(0.0507)	(0.0379)	(0.0481)	(0.0437)
Malicious	0.8503	0.5326	1.9953	1.2124	1.6197	1.8089*
	(0.8921)	(1.5483)	(1.2349)	(0.9240)	(1.1728)	(1.0639)
Negligent	1.5230	0.7308	1.7456	1.0309	1.7072	1.5512
	(0.9960)	(1.7288)	(1.3789)	(1.0318)	(1.3095)	(1.1879)
Processing	-0.2077	0.9619	1.5916	0.8451	3.7457***	1.7439
	(1.0064)	(1.7467)	(1.3932)	(1.0424)	(1.3231)	(1.2002)
Financial	-0.6805**	0.1292	0.1846	-0.1498	0.5426**	0.1091
	(0.3165)	(0.5494)	(0.4382)	(0.3279)	(0.4162)	(0.3775)
Health	-1.3724***	-0.2626	-0.2046	-1.5928***	-2.5531***	-0.7584
	(0.5095)	(0.8843)	(0.7053)	(0.5277)	(0.6698)	(0.6076)
Retail	0.4302	-0.0589	0.1728	-0.0590	0.2895	-0.0136
	(0.3785)	(0.6569)	(0.5239)	(0.3920)	(0.4976)	(0.4514)
Manufac	-1.9037***	-1.0644	-1.5019**	-0.6030	-0.4622	-0.9391*
	(0.4463)	(0.7747)	(0.6179)	(0.4623)	(0.5868)	(0.5322)
Information	0.4550	-0.5595	0.4123	0.0595	1.4463***	0.0848
	(0.3846)	(0.6676)	(0.5324)	(0.3984)	(0.5057)	(0.4587)
Admin	-2.0065***	-0.6310	0.0177	-0.0531	-0.4165	-0.3416
	(0.4343)	(0.7538)	(0.6013)	(0.4499)	(0.5710)	(0.5180)
Yr2014	1.3879***	1.0149**	1.0817**	0.8532***	0.5985*	1.0004***
	(0.2408)	(0.4179)	(0.3333)	(0.2494)	(0.3165)	(0.2871)
Relative security	0.5739**	-0.1530	-0.0165	0.3492	1.0090***	0.2614
	(0.2486)	(0.4315)	(0.9617)	(0.2575)	(0.3268)	(0.2965)
Constant	7.6132***	8.0732***	8.1460***	9.8441***	11.3303***	8.4414***
	(1.0881)	(1.8887)	(1.5064)	(1.1272)	(1.4306)	(1.2978)

Note: This table shows the results of our baseline quantile regression on the response variable, the total costs by cyber risk events. We present estimates and standard errors (in parentheses) of five quantiles (QT) and the OLS model. The independent variables considered in the model are the revenue, three risk types, affiliation to six different industry sectors, a 2014 time break, and the firm's relative security level. *, **, and *** indicate that the p-value is less than the significance levels, 10%, 5%, and 1%, respectively, and coefficients without any indicator show no statistical significance. We use SAS to implement the quantile regression models; the number of observations is 324 after excluding missing values.

Table 8
Quantile Regression Results for Claims Cost Calculation (with the Financial Damage as the Response Variable).

	Response variable = ln(financial damage)					
	QT=0.05	QT=0.25	QT=0.50	QT=0.75	QT=0.95	OLS
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
ln(records)	0.1509***	0.3736***	0.2915***	0.2081***	0.1287***	0.2557***
	(0.0483)	(0.0487)	(0.0361)	(0.0295)	(0.0280)	(0.0347)
ln(revenue)	-0.0464	0.0576	0.0369	0.0789*	0.1549***	0.0753
	(0.0686)	(0.0690)	(0.0512)	(0.0418)	(0.0397)	(0.0493)
Malicious	0.1309	-0.2824	1.2968	1.4012	1.4837	1.4488
	(1.6704)	(1.6816)	(1.2482)	(1.0192)	(0.9674)	(1.2004)
Negligent	-1.1470	-1.9059	-1.0456	0.5098	0.6485	-0.1012
	(1.8650)	(1.8777)	(1.3937)	(1.1380)	(1.0802)	(1.3403)
Processing	2.4134	0.6745	0.9350	1.3945	3.4549***	1.8315
	(1.8844)	(1.8971)	(1.4082)	(1.1498)	(1.0914)	(1.3542)
Financial	0.3341	0.3700	0.5167	0.0823	0.1563	0.2104
	(0.5927)	(0.5967)	(0.4429)	(0.3616)	(0.3433)	(0.4259)
Health	0.6399	-0.0692	0.3997	-1.0097*	-2.2670***	-0.1566
	(0.9540)	(0.9604)	(0.7129)	(0.5821)	(0.5525)	(0.6856)
Retail	0.1494	0.2367	0.5377	-0.0759	0.0923	-0.0702
	(0.7087)	(0.7135)	(0.5296)	(0.4324)	(0.4104)	(0.5093)
Manufac	-1.2273	-1.4290*	-1.3618**	-0.0668	-0.1504	-0.7739
	(0.8357)	(0.8413)	(0.6245)	(0.5099)	(0.4840)	(0.6006)
Information	-0.6795	-1.4039*	0.9969*	0.4620	1.4488***	-0.1093
	(0.7202)	(0.7250)	(0.5382)	(0.4394)	(0.4171)	(0.5175)
Admin	-1.0403	-1.6976**	-0.6254	0.2738	0.3262	-0.5198
	(0.8132)	(0.8187)	(0.6077)	(0.4962)	(0.4710)	(0.5844)
Yr2014	1.9039***	1.0305**	1.0080***	0.8732***	0.6438**	1.0146***
	(0.4508)	(0.4539)	(0.3369)	(0.2751)	(0.2611)	(0.3240)
Relative security	-0.2552	-0.3431	-0.2307	0.4574	0.6766**	0.0386
	(0.4655)	(0.4686)	(0.5077)	(0.2840)	(0.2696)	(0.3345)
Constant	7.8984***	7.5596***	8.5745***	9.7490***	10.2088***	7.8180***
	(2.0375)	(2.0513)	(1.5226)	(1.2432)	(1.1800)	(1.4642)

Note: This table shows the results of our baseline quantile regression on the response variable, the total costs by cyber risk events. We present estimates and standard errors (in parentheses) of five quantiles (QT) and the OLS model. The independent variables considered in the model are the revenue, three risk types, affiliation to six different industry sectors, a 2014 time break, and the firm's relative security level. *, **, and *** indicate that the p-value is less than the significance levels, 10%, 5%, and 1%, respectively, and coefficients without any indicator show no statistical significance. We use SAS to implement the quantile regression models; the number of observations is 324 after excluding missing values.

in today's digitally connected environment, firms can only offer reliable financial reports with secure computer-based information systems. Large firms tend to see cybersecurity enhancement as a significant element of their internal controls for financial reporting systems. Additionally, the literature has documented scale effects in the compliance of reporting requirements to the disadvantage of SMEs (Keasey and Short, 1990).[25]

We also identify that the relative security level is statistically significant and positive at both the extreme lower and upper quantiles in explaining the total cost of cyber events. Although the results in Table 6 show an insignificant coefficient estimate of this variable at the 95% quantile, we find that the estimates between the 80% and 90% quantiles are positive and significant at the 5% or 1% significance level.[26] This tendency is also identified in Table 7, where the number of breached records variable is considered. The results indicate that a relatively weak cybersecurity system is likely to lead to high total costs for firms in the extreme lower or upper region of the cost distribution.[27]

A possible explanation for the result in the extreme upper region might be the results of Zhao, Xue, and Whinston (2013), which indicate that hackers can evaluate the security level of firms with progressively advanced techniques. Hence, the enhancement of security systems of other firms with relatively stronger interconnections (i.e., in the same industry) could lead to negative externality, affecting one's loss probability (the one with weaker security) and thus the size of the loss.[28] Considering the result in the extreme lower region, it is evident that events with marginal costs may frequently occur for firms with relatively weaker security. Our raw data indicate that most marginal-sized events are caused by privacy breach or identity theft (86.9% out of 61 cases below $1,000 cost), which can happen more frequently than other types (e.g., malicious hacking/data breach or network disruption, which leads to much larger loss events).

Concerning the type of risk, we find no statistical evidence of the impacts of malicious or negligent types. However, as shown in Tables 6, 7, and 8, it is commonly observed that firms in a huge cost area (i.e., the 95% quantile) are more likely to be affected by network disruption/processing errors. The network disruption/processing error risks in this dataset can address business interruptions in the interconnected network environment, generating high indirect costs to victims. Business interruption by cyber risks tends to cause extreme losses and impacts with physical (interruption of operations) and non-physical damages (contingent business inter-

[25] Empirical results also show that many small firms tend to take a "wait-and-see" strategy toward cybersecurity investment (Gordon et al., 2018). This tendency is, for example, observed in the UK government's annual survey for cybersecurity breaches, where larger firms tend to see cybersecurity as a high priority than small firms (DCMS, 2021). The underinvestment in cybersecurity might lead to larger losses.

[26] The results of other quantiles are available upon request.

[27] The results are not as strong as one might expect for the other quantiles of the cost distribution. A possible explanation is that security enhancement as self-protection primarily affects loss probability but not the cost. Our results claim that, for the extreme lower or upper-cost events, weaker security would lead to the cost amount, whereas it would not affect the cost amount for the central parts of the cost distribution. To further validate the role of the security level concerning the costs of cyber events, other proxies might be needed for future research.

[28] Zhao et al. (2013) provide a relevant argument to this finding of the impact of other firms' security that a firm's self-protection on its security system can divert hackers (i.e., those who can make strategic decisions) to other firms, thereby increasing other firms' risks. They refer to this consequence as a negative externality (in particular, resulting from one's security investments) that Muermann and Kunreuther (2008) also address with the Nash equilibrium.

ruption, e.g., reimbursement of lost profits) (OECD, 2017). Thus, business interruption could make it more challenging for cyber insurers to adequately estimate risk premiums.

The results in Table 6 also show that, unlike our expectations, industry coefficients generally are not statistically significant, indicating no clear relationship between industry affiliation and total costs. However, the results provide some evidence that the coefficients of the financial and health care/medical industries are negative and significant at the 5% and 95% quantiles, respectively. Additionally, the estimate of the financial industry at the 95% quantile is positive despite the p-value of approximately 10%, and this variable is positive and significant in Table 7. These findings imply that firms in the upper areas of cost distribution are less likely to be in the health/medical care industries; firms in the lower (upper) cost areas are less (more) likely to be in the financial industry.

The results also illustrate that events after 2014 are more likely to significantly impact an increase in the economic cost across the cost distribution. This finding is important for cyber insurers—the economic costs of cyber risk events tend to increase significantly, which can potentially worsen the profitability of the overall market with increasing claims. The results imply that the potential premium risk for insurers against SMEs is expected to increase. Therefore, a pricing adjustment is necessary.

4.2. Claims cost calculation of data breach loss events

Table 7 demonstrates that the number of breached records coefficients are statistically significant and positive across all quantiles at the 1% significance level. The results indicate that total costs generally increase as the number of breached records increases. Further, notably, the impact of a breached record on the economic cost is higher for a smaller loss event, whereas its impact is smallest at the most extreme upper quantile (95%). We compare the magnitude of coefficient estimates, for example, between the 5% quantile and the 95% quantile. Subsequently, we find that the impact of the breach size at the 5% quantile is 52.9% higher than that at the 95% quantile. This magnitude can be interpreted as a cost factor per breached record, which accounts for the increase in cost as the breach size increases. In Fig. 2, we plot the relationship between the firm size and the cost per record, which supports evidence of decreasing the cost per record as the firm size decreases.

As indicated in Section 4.1., our raw data show that firms in the lower economic cost quantiles tend to be SMEs, whereas those in the higher economic cost quantiles are relatively high. Thus, our empirical finding demonstrates that small firms are more likely to be exposed to higher costs per record than large firms. High cost per breached record could cause a severe problem for SMEs increasingly exposed to cyberattacks, with 43% of cyberattacks targeting small businesses (Steinberg, 2019); more SMEs have reported their experiences of cyber incidents (Hiscox, 2019; Ponemon Institute, 2019b).[29]

As the cyber-insurance market has grown significantly, SMEs exposed to greater risk with a relatively weaker security system are expected to give more consideration to a risk transfer option because the premium can be cost-efficient in their risk management against cyber events.[30] However, according to A.M. Best

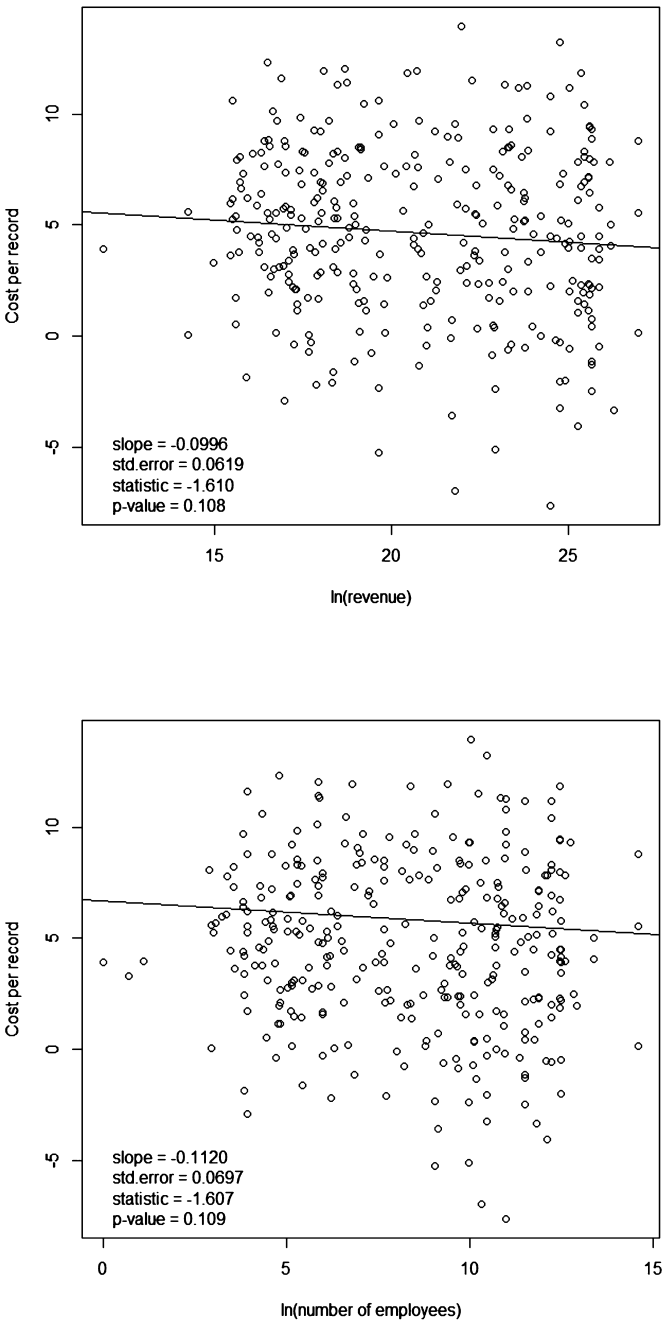

Fig. 2. The relationship between cost per record (log-transformed total costs divided by the number of breached records) and firm size on the x-axis (log-transformed revenue on the left panel and log-transformed number of employees on the right panel). Both have similar sizes of slopes, representing decreasing cost per record over firm size. The slopes are nearly statistically significant, with p-values of 10.8% and 10.9%. The slopes address that a 1% increase in the annual revenue and the number of employees can lead to a 10% and 11% decrease, respectively, in cost per record.

[29] Notably, the result in Table 7 requires a different aspect to interpret it from that of Table 6 because the model in Table 7 discusses how to calculate claims when a loss occurs. However, the model in Table 6 examines what determines the total cost of cyber events in discussing insurance pricing. Thus, the discussion of cost per record only applies to the claims cost calculation in Table 7, which can lead to an argument that the small-sized loss events are more likely to happen to SMEs and may result in larger claims than a cyber-insurer predicted.

[30] A.M. Best (2019) uses the Cybersecurity and Identity Theft Insurance Coverage Supplement data of the National Association of Insurance Commissioners (NAIC) to study the status-quo of the United States cyber-insurance market. A.M. Best (2019) posits that, although SMEs are at greater risk with weaker cybersecurity, the market for SMEs has been underdeveloped until recently. Nevertheless, insurers see ample room for this market to expand significantly; therefore, they will continue to participate in this market. This prediction can also be supported by the status-quo of the market with high profitability (A.M. Best, 2019; Xie et al., 2020), possibly increasing competition in the market. Thus, there exists a high expectation of the development and premium growth of the SME market. Additional evidence of this is provided by Munich Re (2020), and it presents an increasing demand of SMEs for cyber-insurance in the firms that are more aware of their exposure.

(2019), it has been observed in the current cyber-insurance market that total cyber claims of SMEs have increased rapidly in recent years, although they have lower limits and fewer protections with commensurate premiums. Indeed, claims are increasing exponentially, which has become an increasing concern for cyber insurers that have taken advantage of the effect of an emerging market with substantial profitability over the last few years (A.M. Best, 2019; Xie et al., 2020).

A potential problem based on the findings of Xie et al. (2020) is that insurers with a few lines of business (i.e., less growth potential as addressed by the authors) are more likely to be in the current cyber-insurance market, representing a high potential risk due to significantly increasing claims. In line with our findings, cyber insurers with a few lines of business might primarily face small-sized events that have a higher cost per record, leading to higher numbers of claims than they predicted. Thus, we claim that cyber insurers should consider the heterogeneous effects of firm size on total costs when establishing a predictive model for potential cyber claims. This consideration leads to the conclusion that insurers should provide heterogeneous premiums for firms of different sizes.[31]

4.3. Alternative measure for the cost of claims and comparison with the OLS fit

We use the financial damage of cyber risk events as an alternative measure of the number of losses to examine the impact of breach size and firm specifics on the loss amount for the claims cost calculation model. This financial damage does not include the litigation cost, which differs from the main target variable (total costs). Financial damage represents the amount of loss driven purely by cyber risk events. Although total costs and financial damage variables have some identical observations, observations of total costs are generally higher due to the inclusion of litigation costs (e.g., legal and defense costs). Thus, we do not include litigation variables in Table 8. This exclusion can help us understand which loss-quantile firms tend to be more affected by litigation.

Table 8 shows quantile regression results using financial damage as the dependent variable, where the number of breached records again is statistically significant at the 1% level across all quantiles. The tendency that estimates at lower quantiles is larger is still present in this case. The magnitudes of the coefficients of breach size estimated from the 25% to 75% quantiles are larger than those in Table 7. In line with the finding in Table 7, the litigation cost may account for a large portion of loss events at the moderate level. Hence, insurers may be more burdened by liability costs for moderate losses that would be above deductibles.

We also compare our quantile regression results with the OLS method in Tables 6, 7, and 8. The traditional least squares regression techniques estimate the average effect of firm-specific and industry factors and may not represent the properties of non-central locations in the cost distribution, providing an incomplete picture of the relationship. We examine the deviation of key variable estimates (revenue in Table 6 and the breached records in Tables 7 and 8) at the 5% and 95% quantiles from the OLS result. We show a significant difference between these two results. This difference implies that focusing on the average effects using the OLS may under- or over-estimate the relevant coefficients or even fail to identify some heterogeneous relationships between the lower and upper parts of the cost distribution.

Further, we identify in Tables 6 and 7 that the OLS approach might result in overlooking the impacts of the affiliation in the healthcare, medical, or financial industries and a firm's relative security level on the determination of the total cost. However, the impacts are significant in explaining the total cost at extreme quantiles. Fig. 3 supports the findings by showing that the slope of the 5% quantile (the very bottom line) is steeper than the others, and extreme quantiles tend to have steeper slopes than less extreme quantiles. The scatter plots in both panels appear to diffuse, implying potential deviations in the impacts of loss drivers that the OLS method with the conditional mean of the response variable might not capture. As the quantile regression captures the properties of particular locations across the different quantile levels of the cost distribution, the relationship between the target variable and the predictors in this diffuse type can be assessed more precisely.

The OLS line (red) is nearly identical to the median line of quantile regression and significantly different at the extreme lower and upper quantiles. Fig. 4 shows the estimates of three key variables—the number of breached records, revenue size, and the number of employees—across quantiles from 5% to 95% compared to the OLS fit (red line) and its 95% confidence interval (dotted lines). As these plots are based on a simple regression, the sizes of the estimates can be slightly different from those of the multivariate setup in Tables 6, 7, and 8. All three variables have particularly significant intervals in coefficient estimation compared to the confidence intervals of the OLS fit. In particular, the number of breached records has more significant impacts at the extremely lower quantiles (5–20%) and the upper quantiles (75–90%).

4.4. Robustness check with the number of employees

As discussed in the previous sections, the size of a firm can play an important role in determining the total costs of cyber risk events. Thus, cyber insurance pricing should be differentiated according to firm size (not linear but non-linearly). To support this proposal, we check the robustness of our findings using the number of employees as an alternative measure of firm size in all the considered models. The results in panels A, B, and C of Table 9 focus on how selected key variables (i.e., the firm size and the number of breached records) are affected by the use of the alternative measure of firm size. This robustness test presents consistent results with the main results (Tables 6, 7, and 8) for the key variables. Specifically, the statistical significance and signs of the firm size (number of employees) and the number of breached records variables are consistent. The tendency for heterogeneity in the impacts of key variables still holds, particularly in Tables 6 (insurance pricing aspect) and 7 (claims cost calculation aspect).

We consider two more models to check the robustness of the findings. One aspect is the exclusion of risk types for insurance pricing. This aspect may hold if one sees risk types as ex-post information; thus, they may not be influential when pricing. To consider this aspect, we exclude three risk-type variables from the model in Table 6. The second aspect is investigating the impact of a firm's revenue on the unit cost per exposure. This aspect helps check our key finding that the impact of a firm's revenue is stronger (weaker) in the lower (upper) levels of total costs. One may expect that the larger a firm is, the higher the cyber risk exposure it may face. However, as explained in Section 4.1, economies of scale for cybersecurity investments, regulatory compliance, and the "wait-and-see" strategy may drive lower unit costs for larger firms. To support this finding, we use the ratio between

[31] Ponemon Institute (2020) determines that the cost of loss events associated with a data breach can be explained by three factors—lost business, detection and escalation, and ex-post response. They surveyed 2,176 individuals from small- and mid-sized companies in 11 countries in 2019. Consequently, they found that SMEs tend to have ineffective IT security postures due to mainly insufficient personnel, budget concerns, and security technologies (Ponemon Institute, 2019b). It can imply with our finding that SMEs might struggle with a higher cost per record with a high fixed cost for recovery, possibly resulting from the lack of an effective strategy to respond to loss events.

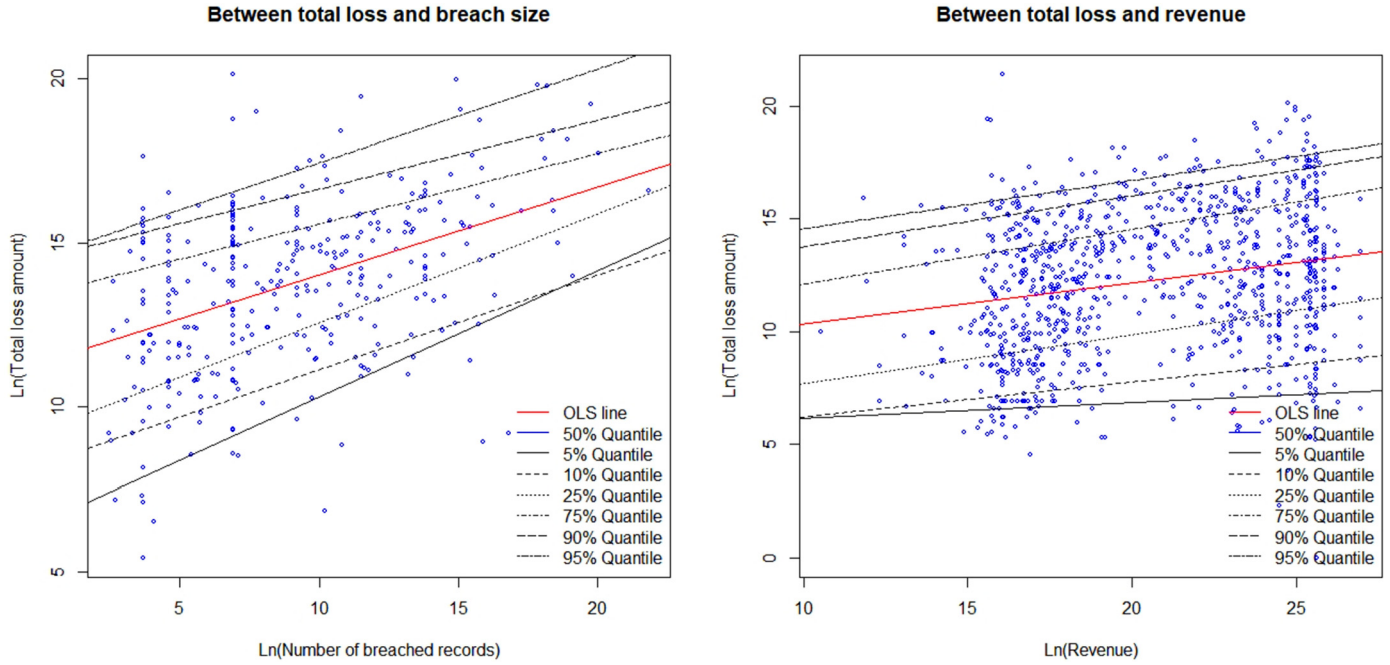

Fig. 3. Comparison of OLS and quantile regression estimates. These plots provide a comparative setting in model fitting between the OLS method and quantile regression with different quantiles. The left panel illustrates the relationship between the total costs and the number of breached records; the right panel shows one between the total costs and the revenue size. In both plots, quantile lines below 50% describe the relationship in small-sized losses; lines above 50% show the relationship in large-sized ones. As expected, the 50% quantile (median) line does not deviate from the OLS line in both plots; other quantile lines seem to be significantly away from the central line.

total costs and a firm's revenue as the response variable for the insurance pricing model in Table 6.

The results of both aspects are presented in panels D and E of Table 9, consistent with Table 6 concerning the key variable. In particular, the model for the second aspect in panel E demonstrates that the cost of a cyber risk event per revenue unit is much lower in the upper quantile, where firm size tends to be large. Thus, we can conclude that the size effect resulting from economies of scale and corporate decision-making on cybersecurity investments is valid in determining the size of cyber risk costs.

5. Applications to cyber-insurance rate-making

5.1. Two-part GLM and Tweedie model for cyber risk

We use quantile regression to show the heterogeneous impacts of the firm- and industry-specific factors on the total costs across different quantiles. Now, we assess how the quantile regression results differ in empirical pricing from standard models by deriving pure premiums based on the estimates obtained by varying the firm's size, industry, and security state. This application has two significant characteristics: 1) it shows how the quantile regression method can be used in pricing cyber insurance; 2) it shows how the estimated premiums with parameters of the quantile regression (e.g., the probability of no cyber claims) differ from the premiums based on the classical actuarial pricing methods.

For this comparison, we estimate the premiums from two other pricing schemes—the GLM of the two-stage method and the Tweedie model—the extensive uses of which have been reported in the study (Haberman and Renshaw, 1996; Garrido et al., 2016 for the GLM; Jørgensen and de Souza, 1994; Shi, 2016 for the Tweedie model). For the sake of completeness, we present a brief discussion of the estimations of the GLM using a two-part approach and the Tweedie model implicitly providing a compound process.

A two-part method for the pure premium calculation comprises regression modeling for frequency and severity, where GLMs for frequency and severity are implemented with discrete link functions and link functions from the exponential family (Hsiao et al., 1990; Frees, 2009, Chapter 16).[32] The use of a link function on the response variable is a key difference between the GLM and OLS, the function that enables one to model a non-Gaussian shape of the response variable with the interaction of explanatory variables. It is beneficial for insurance data that typically feature right-skewness in severity and discreteness in frequency.

Insurance claims data generally comprise several zeros and non-zeros reported as claims. We denote the occurrence of a claim by r_i $(i = 1, \ldots, N)$, with the insurance portfolio size of N, and the size of a claim by y_i; r_i is a binary variable defined as follows:

$$r_i = \begin{cases} 0, & \text{if } y_i = 0 \\ 1, & \text{if } y_i > 0 \end{cases} \tag{5}$$

When the claim is observed, the indicator variable, r_i, offers 1 and 0 otherwise. Subsequently, the final claims recorded in the portfolio can be defined as follows:

$$y_i^* = \begin{cases} 0, & \text{if } r_i = 0 \\ r_i \times y_i, & \text{if } r_i = 1 \end{cases} \tag{6}$$

If the frequency exceeds one such that the number of claims includes more than one claim from the i^{th} policy, then the frequency problem would model the number of claims with claims arrival processes (e.g., Poisson or negative binomial) beyond modeling the occurrence of claims. Our dataset provides no information to identify the exact breached entities; thus, we focus on the frequency model with loss occurrence. The loss frequency and severity are modeled separately by examining how the covariates explain each component. Covariates comprise policyholders' characteristics that, in our case, are firm-specific factors used in the quantile regression.

[32] In this application, we take a two-part GLM approach with a gamma function as part of the exponential family for loss severity and a dichotomous claim occurrence for loss frequency (i.e., occurrence or non-occurrence of an event) to accommodate our data with numerous zeros (events with no claim/loss) (Frees, 2009, p. 417).

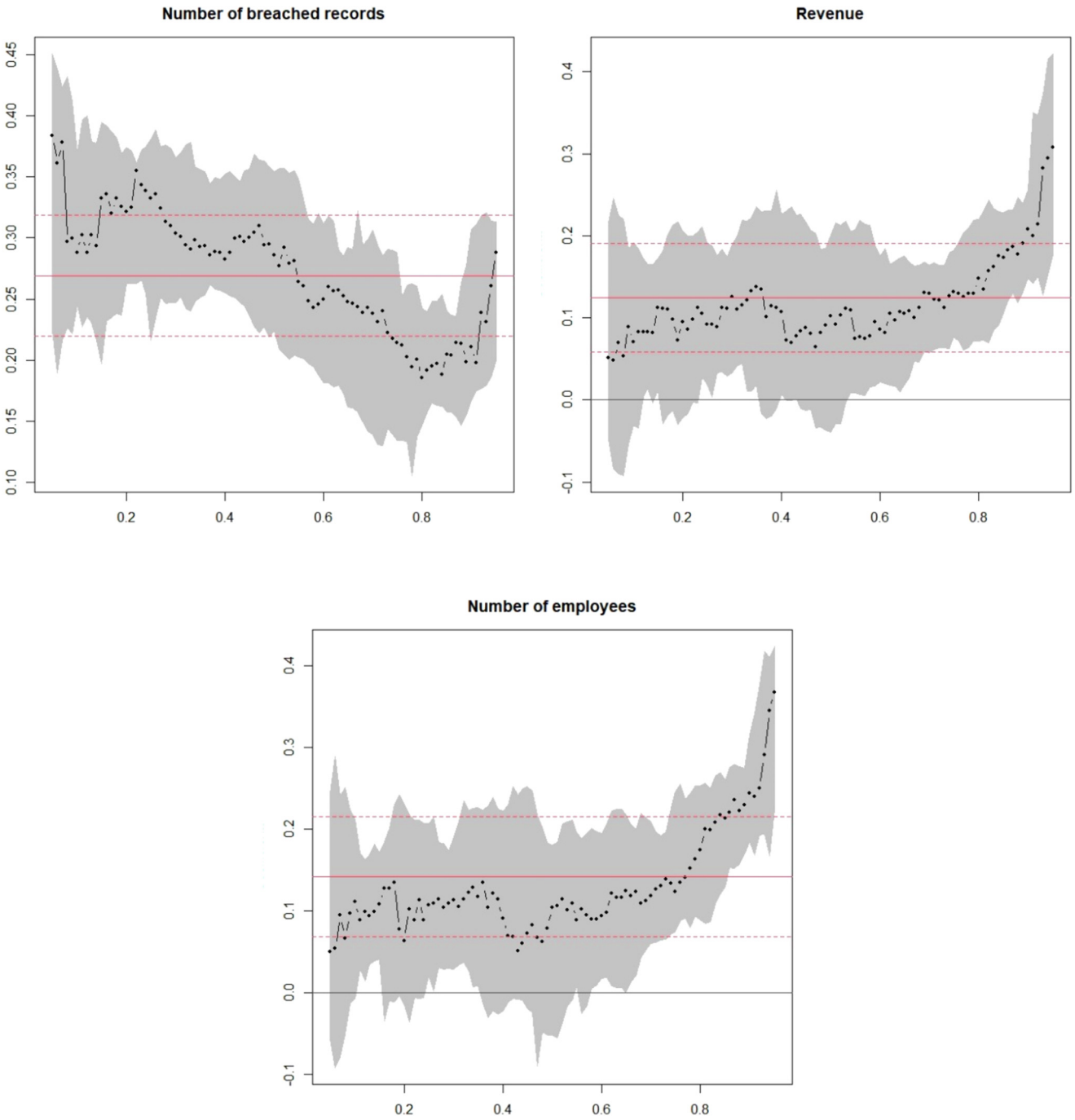

Fig. 4. Changes in estimates of three key variables over quantiles. This graph shows coefficients of three key variables (number of breached records, revenue, and number of employees) over quantiles from 5% to 95%. Coefficients are projected on the y-axis, and quantiles are on the x-axis. The shadowy area indicates 95% confidence intervals of coefficient estimators. The horizontal line between two dotted lines represents the OLS fit (static across quantiles), and two dotted lines are its corresponding 95% confidence interval.

Table 9
Robustness Check Test Results.

Panel A: Robustness check for insurance pricing					
	Response variable = ln(total costs)				
	QT=0.05 Estimate	QT=0.25 Estimate	QT=0.50 Estimate	QT=0.75 Estimate	QT=0.95 Estimate
ln(records)	-	-	-	-	-
ln(employee)	0.0673	0.2175***	0.2084***	0.1781***	0.1671***
	(0.0502)	(0.0407)	(0.0414)	(0.0376)	(0.0312)
Control variables	Yes	Yes	Yes	Yes	Yes
Panel B: Robustness check for claims cost calculation					
	Response variable = ln(total costs)				
ln(records)	0.2238***	0.3051***	0.2653***	0.2064***	0.1471***
	(0.0238)	(0.0367)	(0.0367)	(0.0283)	(0.0333)
ln(employee)	-0.0247	0.1147*	0.1188**	0.1258***	0.1003*
	(0.0381)	(0.0588)	(0.0588)	(0.0454)	(0.0533)
Control variables	Yes	Yes	Yes	Yes	Yes
Panel C: Robustness check for the variable of financial damage					
	Response variable = ln(financial damage)				
ln(records)	0.1906**	0.3557***	0.2901***	0.1910***	0.1263***
	(0.0344)	(0.0499)	(0.0359)	(0.0291)	(0.0293)
ln(employee)	0.1639***	0.0828	0.0897	0.1201***	0.1717***
	(0.0524)	(0.0762)	(0.0548)	(0.0445)	(0.0447)
Control variables	Yes	Yes	Yes	Yes	Yes
Panel D: Robustness check for insurance pricing without risk type variables					
	Response variable = ln(total costs)				
ln(revenue)	0.0823*	0.1902***	0.2111***	0.1957***	0.1766***
	(0.0422)	(0.0399)	(0.0410)	(0.0342)	(0.0387)
Control variables	Yes	Yes	Yes	Yes	Yes
Panel E: Robustness check for the variable of the ratio between total costs and revenue					
	Response variable = (total costs / revenue) × 100				
ln(revenue)	-0.0005***	-0.0056***	-0.0416***	-0.3007***	-5.5122***
	(0.0000)	(0.0003)	(0.0017)	(0.0050)	(0.0101)
Control variables	Yes	Yes	Yes	Yes	Yes

Note: This table shows the results of models to check the robustness of the findings. We present estimates and standard errors (in parentheses) of five quantiles. The table illustrates five model settings, where the first three panels represent the models of Tables 6 to 8 by replacing the revenue with the number of employees, and the last two panels address the model of Table 6 without risk types and using the unit cost as the response, respectively. The results focus on 1) whether statistical significance and heterogeneity in impacts of firm size still hold (panels A, D, and E) and 2) whether those of the breach size remain similar (panels B and C). "QT" and "employee" stand for quantile and the number of employees, respectively. *,**, and *** indicate that the p-value is less than the significance levels, 10%, 5%, and 1%, respectively, and coefficients without any indicator show no statistical significance. We use SAS to implement the quantile regression models in this table.

We use a binomial function as a link to represent the distribution of the occurrence of losses, which is defined as follows (Frees, 2009, Ch. 16.4):

$$r_i = \mathbb{x}_{1i}'\beta_1 + \eta_{1i},$$
$$r_i \sim B(1, \tau), \tag{7}$$

where $\mathbb{x}_{1i}$ is the vector of explanatory variables, β_1 is the vector of coefficients, η_{1i} is the error term with the assumption of i.i.d., and τ is the probability of success (loss probability in this case) of an event.

For the loss severity model, a gamma function is used on the response variable, which is defined as follows (Frees, 2009, Ch. 16.4):

$$y_i = \mathbb{x}_{2i}'\beta_2 + \eta_{2i}, y_i \sim G(\mu, \nu),$$
$$G(y) = \frac{y^{-1}}{\Gamma(\nu)}\left(\frac{y\nu}{\mu}\right)^{\nu} e^{-\frac{y\nu}{\mu}}, \tag{8}$$

where $\mathbb{x}_{2i}$ is the vector of the explanatory variables, β_2 is the vector of the coefficients, η_{2i} is the error term with the assumption of i.i.d., and μ and ν are the shape and scale parameters, respectively, of the gamma distribution for the response variable.

The Tweedie model is an efficient tool that can replace a two-part model splitting loss frequency and severity by providing statistical features of two components in a single model (Tweedie, 1984; Jørgensen and de Souza, 1994). The probability density function of a Tweedie process shows most of its mass at zero, and the rest is highly skewed to the right. Thus, it applies to a dataset in which numerous zeros and highly skewed non-zeros are mixed. As illustrated in Table 10, the well-known discrete and continuous distributions used in the insurance literature can be represented by the Tweedie model as per its power parameter, $p \in \mathrm{R}$ (Ohlsson and Johansson, 2010, Chapter 2). For more details on the specification of the Tweedie distribution, see Jørgensen and de Souza (1994), where the mean and variance parameters and their relationships with the power parameter are explained in more detail.

A Tweedie process with a power parameter between 1 and 2 ($p \in (1, 2)$) is generally defined as a Poisson sum of a gamma dis-

Table 10
Distribution Specification of Tweedie Models.

Power parameter	Distribution type	Specific family
$p < 0$	Continuous	-
$p = 0$	Continuous	Gaussian
$0 < p < 1$	Non-existing	-
$p = 1$	Discrete	Poisson
$1 < p < 2$	Mixed, non-negative	Poisson-Gamma (Compound)
$p = 2$	Continuous	Gamma
$2 < p < 3$	Continuous	-
$p = 3$	Continuous	Inverse Gaussian
$p > 3$	Continuous	-

tribution, representing a combination of Poisson and gamma distributions. If a dataset follows a Tweedie process with the power parameter in this range, a two-part approach would not be needed, which offers a convenient and efficient model in insurance pricing. We estimate the coefficients of the predictors (explanatory variables) for loss frequency with logistic regression (as part of the GLM) and severity with a log link of the gamma GLM model. The specifications of the two models are as follows:

Loss frequency:

$$h(\mu_1) = \beta_{1,0} + \beta_{1,1}x_{1,1} + \beta_{1,2}x_{1,2} + \cdots + \beta_{1,11}x_{1,11},$$
$$h(\mu_1) = \ln\frac{\mu_1}{1-\mu_1}, \tag{9}$$

Loss severity:

$$g(\mu_2) = \beta_{2,0} + \beta_{2,1}x_{2,1} + \beta_{2,2}x_{2,2} + \cdots + \beta_{2,11}x_{2,11},$$
$$g(\mu_2) = \log(\mu_2), \tag{10}$$

where μ_1 and μ_2 are the means of the binomial and gamma distributions, respectively.

We use the maximum likelihood estimation method to determine the optimal power parameter of the Tweedie model, which shows a compound process as expected in a typical insurance data ($p = 1.7$). Table 11 shows the estimation results of a two-part GLM and the Tweedie model. Notably, many variables have statistically significant impacts on the loss occurrence (frequency), whereas this is not the case for loss severity. This result is expected in loss/cost analysis because variables tend to be useful in predicting frequency but not as useful in explaining severity (Frees, 2009, p. 372). For the Tweedie model, only the retail variables are statistically significant.

The most important points in the GLM and Tweedie approaches are building the model and selecting the variables for insurance rate-making. The selection of the important determinants for the cyber claims process is slightly different from that of the quantile regression, including the median model ($q = 0.5$), which is comparable with the GLM and Tweedie models because neither of the models differentiates quantile effects. The following section examines how this difference between the quantile regression results and the two models' results reflect the heterogeneous impacts of price determinants in insurance premiums.

5.2. Comparison in cyber-insurance rate-making

Now, we compare the predictive models with the quantile regression, the two-part GLM, and the Tweedie model, which we use to calculate the potential premiums of cyber insurance. As mentioned in the previous section, the extant literature provides an insurance rate-making scheme with quantile regression, referred to as the *quantile premium principle* (Kudryavtsev, 2009; Heras et al., 2018; Baione and Biancalana, 2019). The common idea of the quantile premium principle proposed in the literature is as follows:

Step 1) Determination of the quantile for pricing. The quantile that should be used for regression is determined by the following equation (Kudryavtsev, 2009; Heras et al., 2018):

$$\tilde{\theta} = \frac{\theta - \pi}{1 - \pi}, \tag{11}$$

where θ is the cumulative probability of the actual loss of a policy not exceeding its premium paid, which can be regarded as the solvency rate (thus, $1-\theta$ can be the ruin probability). Further, π is the probability that a policyholder makes no claim, and $\tilde{\theta}$ is the risk-based quantile estimate used for the quantile premium principle.

Step 2) Establishment of the quantile regression model. Equation (11) with the combination of θ and π leads to the modified risk-based quantile ($=\tilde{\theta}$) to construct a pricing model. The model with the modified quantile is defined based on Equation (11) as

$$Q_{\tilde{\theta}}(y_i \mid x_i) = x_i'\beta_{\tilde{\theta}} \tag{12}$$

Step 3) Pure premium calculation with the estimates of the quantile model. The optimal quantile for pricing might be different for a given risk class (e.g., firm size or relative security level in the cyber-insurance case).

The estimated quantile $\tilde{\theta}$ is the key to this quantile-based insurance rate-making. This quantile estimator accommodates material features of rate-making by considering 1) the distributional property of losses and 2) the information on policies without losses (Kudryavtsev, 2009). Hence, it facilitates risk-based premium calculation by enabling loss frequency and severity to be inherent in the pure premiums. Table 12 shows the pure premiums based on quantile regression and those from the gamma GLM and Tweedie models. To clarify the comparison, we use three firms' sizes (large, medium, and small) represented by either the annual revenue (panel A) or the number of employees (panel B) from the financial and health care/medical industries.[33] We consider whether a potential policyholder has relatively weak security. We estimate the premiums for hypothetical cyber insurance policies against malicious risk attacks, categorized in Table 3. The key used to differentiate the hypothetical insureds is the firm size and the relative security level, which are the primary factors in underwriting cyber insurance.[34]

For the quantile premium principle, the modified quantile is dependent on both the probability of claims ($= 1-\pi$) and the quantile to be set (θ), which can vary according to the firm's characteristics. In our dataset, $1-\pi$ turns out to be 4.91% for the financial industry and 2.93% for the medical industry. These numbers are calculated by the ratio between the number of claims reported (i.e., non-zero total costs) and the total observations in the entire dataset as the size of the cyber-insurance portfolio. This calculation provides a proxy to measure the number of claims occurring in the cyber-insurance portfolio and is in line with the probability of the claims used in other studies (Eling and Schnell, 2020). We use these probabilities to determine the quantiles in pricing for medium-sized firms of two industries. For small- and large-sized firms, we diversify the probabilities of claims with a 1% margin from those of medium-sized firms.[35] We use 99.5% of the

[33] We select the financial and medical industries for the applications based on the fact that these two industries have consistently faced high severity/high frequency of events due to malicious attacks as reported by Ponemon Institute (2020). Thus, they are more likely to need financial protection against potential cyber risk events.

[34] There have been practical observations and analyses about how cyber-insurance can be priced from academia and industry. For instance, Franke (2017) analyzes the Swedish cyber-insurance market and finds that pricing is primarily based on expert models rather than a historical database. These expert models use the company size, industry, and the assessment of the security of the company's IT/information as important factors for pricing. AIR Worldwide (2017) also posits that the firm's size is the first characteristic (in terms of revenue) to represent the nature of an insured's cyber risk. A recent study by Zeller and Scherer (2021) considers, among others, types of industry and firm size as key characteristics of firms in explaining their vulnerabilities against cyberattacks. They use those factors to develop their cyber actuarial model for insurance pricing and risk measurement.

[35] This adjustment in the probability of claims for firms of different sizes leads us to estimate quantile premiums at 87.2%, 89.8%, and 91.5% quantiles for large, medium, and small financial firms, respectively, and at 74.1%, 82.9%, and 87.3% quantiles for large, medium, and small medical firms, respectively.

Table 11
Generalized Linear Model and Tweedie for Frequency (binary) and Severity.

	Firm size with revenue						Firm size with the number of employees					
	Generalized Linear Model				Tweedie		Generalized Linear Model				Tweedie	
	Frequency (logistic)		Severity (Gamma)		$p^* = 1.7$		Frequency (logistic)		Severity (Gamma)		$p^* = 1.7$	
	Estimate	Test-statistic	Estimate	Test-statistic	Estimate	Test-statistic	Estimate	Test-statistic	Estimate	Test-statistic	Estimate	Test-statistic
ln(revenue)	0.0767	4.942***	0.0841	1.560	0.0609	1.003						
	(0.0155)		(0.0539)		(0.0608)							
ln(emp)							0.0815	4.675***	0.1086	1.828*	0.0695	1.030
							(0.0174)		(0.0594)		(0.0675)	
Malicious	-4.1367	-5.635***	0.0500	0.061	0.0537	0.059	-4.1397	-5.641***	-0.0133	-0.016	0.0165	0.018
	(0.7341)		(0.8164)		(0.9105)		(0.7339)		(0.8109)		(0.9137)	
Negligent	-3.0452	-3.823***	-0.4918	-0.411	-0.8871	-0.602	-3.0328	-3.809***	-0.4989	-0.420	-0.9068	-0.614
	(0.7965)		(1.1963)		(1.4729)		(0.7963)		(1.1871)		(1.4769)	
Processing	-0.6663	-0.744	1.1928	1.238	1.1194	1.087	-0.6593	-0.736	1.1206	1.172	1.0925	1.057
	(0.8960)		(0.9637)		(1.0302)		(0.8956)		(0.9565)		(1.0332)	
Financial	-0.4122	-2.936***	0.2873	0.575	0.1492	0.252	-0.3458	-2.531**	0.3198	0.665	0.1982	0.344
	(0.1404)		(0.4999)		(0.5911)		(0.1366)		(0.4807)		(0.5761)	
Health	0.1321	0.472	-1.2959	-1.450	-1.1575	-0.942	0.1258	0.450	-1.3020	-1.469	-1.1553	-0.937
	(0.2796)		(0.8934)		(1.2290)		(0.2795)		(0.8866)		(1.2336)	
Retail	-0.2569	-1.471	2.4339	3.946***	1.8711	2.930**	-0.2268	-1.306	2.5025	4.124***	1.8976	2.991**
	(0.1746)		(0.6169)		(0.6385)		(0.1737)		(0.6069)		(0.6345)	
Manufac	0.0144	0.059	0.5097	0.668	0.6159	0.733	0.0846	0.352	0.5570	0.738	0.6507	0.774
	(0.2422)		(0.7630)		(0.8409)		(0.2405)		(0.7545)		(0.8410)	
Information	0.1312	0.784	1.0240	1.886*	0.7055	1.149	0.1753	1.057	1.0061	1.887*	0.7332	1.203
	(0.1674)		(0.5430)		(0.6140)		(0.1658)		(0.5331)		(0.6096)	
Admin	-1.8921	-11.99***	0.5382	0.821	0.3560	0.452	-1.9131	-12.11***	0.5434	0.835	0.3478	0.441
	(0.1578)		(0.6558)		(0.7873)		(0.1580)		(0.6507)		(0.7893)	
Yr2014	-1.2455	-12.84***	0.6889	1.849*	0.4918	1.230	-1.2393	-12.78**	0.6782	1.834*	0.4908	1.222
	(0.0970)		(0.3725)		(0.3999)		(0.0970)		(0.3698)		(0.4017)	
Relative security	-0.0547	-0.514	0.3513	0.917	0.2042	0.477	-0.0460	-0.431	0.3190	0.839	0.2065	0.482
	(0.1065)		(0.3830)		(0.4280)		(0.1067)		(0.3801)		(0.4282)	
Constant	2.6758	3.435***	12.9664	10.35***	13.7794	9.748***	3.5533	4.789***	13.8709	15.17***	14.4782	13.88***
	(0.7791)		(1.2531)		(1.4136)		(0.7420)		(0.9145)		(1.0435)	

Note: This table shows the results of two approaches: Generalized Linear Model and Tweedie. GLM includes two-part modeling for frequency (binary) and severity, where loss frequency is modeled with logistic regression, and severity is with gamma link function. The optimal power parameter for the Tweedie model (= p^*) turns out to be 1.7, indicating a Poisson-distributed sum of gamma distribution as a compound process. The independent variables considered in the model are the revenue (or the number of employees), three risk types, affiliation to six different industry sectors, a 2014 time break, and the firm's relative security level. "QT" and "emp" stand for quantile and the number of employees, respectively. The figures in the parentheses of estimate cells are standard errors. *,**, and *** indicate that the p-value is less than the significance levels, 10%, 5%, and 1%, respectively, and coefficients without any indicator show no statistical significance. We use the *R* package, Tweedie, and cplm for the Tweedie modeling and the base package for the gamma GLM.

Table 12
Comparison in Pure Premium of Cyber-insurance by Varying the Firm Size.

Panel A: Firm size of revenue

	Insured A				Insured B				Insured C			
Firm size	Large (Revenue: $20 billion)				Medium (Revenue: $0.2 billion)				Small (Revenue: $0.02 billion)			
Industry	Financial		Medical		Financial		Medical		Financial		Medical	
Security	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec
Quantile	25,081.1	14,949.0	5,723.8	4,051.7	20,533.8	10,226.9	2,791.4	1,453.3	13,237.5	7,503.1	1,892.7	1,116.1
GLM	10,611.6	7,383.1	1,896.9	1,312.3	7,674.3	5,354.9	1,418.0	984.8	6,485.0	4,530.5	1,215.4	845.5
Tweedie	10,773.6	8,783.8	2,916.6	2,377.9	8,137.9	6,634.9	2,203.0	1,796.2	7,072.7	5,766.5	1,914.7	1,561.1

Panel B: Firm size with the number of employees

Firm size	Large (Num. employees: 23,000)				Medium (Num. employees: 1,300)				Small (Num. employees: 180)			
Industry	Financial		Medical		Financial		Medical		Financial		Medical	
Security	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec	Weaksec	Strsec
Quantile	27,181.9	14,546.4	3,286.8	1,817.3	17,652.3	9,818.6	3,281.5	1,515.3	12,088.8	8,101.6	1,806.0	966.5
GLM	9,960.6	7,173.7	1,758.0	1,261.1	7,604.6	5,485.7	1,367.6	983.1	6,287.8	4,540.2	1,143.9	823.3
Tweedie	10,405.5	8,464.0	2,688.1	2,186.6	8,521.3	6,931.4	2,201.4	1,790.7	7,426.9	6,041.2	1,918.6	1,560.7

Note: Insured A, B, and C are hypothetical corporations in two industries (financial and health care/medical) demanding insurance policies against malicious cyber risks. We refer to descriptive statistics for insureds' profiles by considering Q3, Q2, and Q1 values of revenue and the number of employees for the firm size. Assuming the solvency rate of firms as 99.5% ($\theta = 0.995$), we take the probability of claims (= $1 - \pi$) of 1) 4.91% for the mid-sized financial firm from the data and assign it for the large-sized and small-sized ones with $\pm 1\%$ and 2) 2.93% for the mid-sized medical firm from the data and assign it for the large- and small-sized ones with $\pm 1\%$. This parameter set-up leads us to use 87.2, 89.8, and 91.5% quantile (= $\tilde{\theta}$) for large, medium, and small financial firms, respectively, and 74.1, 82.9, and 87.3% quantile for medical firms, for quantile premium pricing. The unit of results is dollar ($).

cumulative probability (= θ) for the actual loss not exceeding the premium paid (i.e., the solvency rate of the insured).[36]

Table 12 indicates that the quantile premiums from our data are generally higher across industry and firm size than those from the other two models. One aspect of this result is that the impact of the firm-size variable in the quantile-based model for our data is much higher than that of the others. In particular, a larger deviation of premiums by the firm's size is more prominent for the quantile-based method than for the others.[37] This result is reasonable as premiums for large firms are estimated at lower quantiles driven by lower probabilities of claims; the models at lower quantiles demonstrate a higher impact of the firm size, as shown in Table 6. Furthermore, this result for a larger deviation by the firm size can help us demonstrate how the heterogeneous impacts of firm size are embedded in the insurance premiums of the quantile model. Therefore, this implies that the impact of the firm size in the quantile-based model is more significant than that in the other models.

Another aspect to explain under- or over-estimation of premiums between classical models and the quantile premium principle (QPP) is the actual probability of claims (= $1 - \pi$) for different insureds. Dependency on the actual probability of claims is associated with the parameter estimation for the quantile in pricing, as shown in Equation (11); hence, the premium estimates depend on the quantile estimates. Fig. 5 addresses this relationship by plotting the premium difference (i.e., the premium of the QPP minus that of the other models) for different quantile estimates in the case of financial firms. It shows that the magnitude of the premium difference increases with the quantile estimate.

We observe that the difference becomes minimal between the 75% and 80% quantiles. This threshold aligns with the relevant studies that also take the quantile estimates within this range (Kudryavtsev, 2009 with 75% quantile estimate; Baione and Biancalana, 2019 with 79.1% quantile estimate). In our case in Table 12, the estimated quantiles are data-driven for each potential insured. The estimates generally range between the 80% and 90% quantiles, resulting in larger premiums for the QPP than for the other models. This also implies that the opposite case can occur if the quantile estimate is lower than a threshold depending on the solvency rate and the claim probability. Thus, it can lead to an essential deviation from other pricing models. Overall, this heterogeneity from the quantile model in pricing (i.e., heterogeneous pricing scheme over the quantile estimate) may help cyber insurers differentiate risks by taking a lower quantile with a higher firm-size effect and a higher quantile with a smaller impact of firm size, based on firms' risk exposures (represented by the likelihood of claims).

Concerning the impact of the relative security level, we find that a weak security level increases risk and, thus, the premium. The impact of the relative security level can almost double the premium of the quantile principle if it is relatively weaker than the industry average. We also find that financial firms are expected to have much larger premiums than medical firms. This result is associated with our finding that financial firms are more likely to experience larger cost events than medical firms. Finally, the results with the number of employees in panel B of Table 12 appear to be in line with those with the annual revenue in panel A in that quantile premiums, generally, are larger in both industries than those of the other models.

6. Conclusion

We propose a quantile-based approach to investigate two aspects of cyber risk research: 1) modeling cyber-insurance pricing and 2) calculation of the total costs by data breach events. The study on firm-specific effects and cost estimation of cyber risk events has focused mainly on either of the two aspects and simple linear relationships. By investigating both aspects with a quantile-based approach, we show that a firm's size is statistically significant in explaining the total costs with heterogeneous effects across different quantiles of the cost distribution. This result suggests that cyber insurers should differentiate insurance premiums based on the firm size, considering the heterogeneity described above.

We also find that firms with security levels weaker than the average for the industry are more likely to be exposed to high-cost

[36] This probability is chosen as the European Solvency II suggests 99.5% to estimate the risk capital, and Baione and Biancalana (2019) adopt this for their quantile premium pricing.

[37] We find that the average deviations in premiums across three firm sizes (when revenue is considered) are approximately 65.2%, 26.6%, and 23.7% for the quantile, GLM, and Tweedie models, respectively. The deviations with the number of employees are 41.0%, 24.9%, and 18.4% for the quantile, GLM, and Tweedie models, respectively.

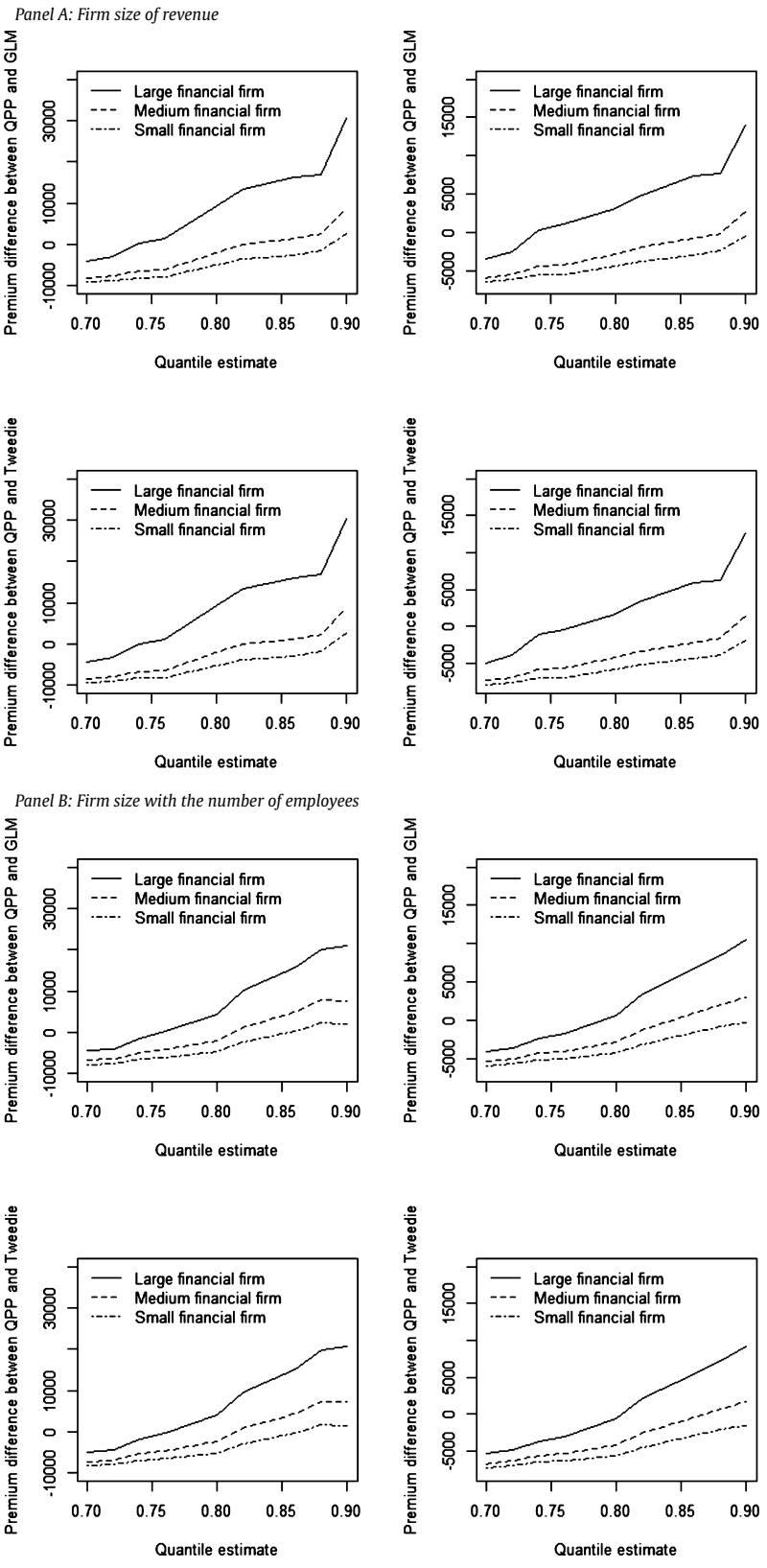

Fig. 5. Differences in premium estimates of financial firms between the quantile pricing scheme and two other methods (dollar unit on the y-axis).

events. This is an important result because practitioners often consider security information the main risk driver. To the best of our knowledge, this has never been examined in an academic study because no firm-specific security information is available. We determine that the risk driven by network interruption/processing errors is more likely to be associated with extreme cost events, in which losses related to business interruption are present. Additionally, the financial (health care/medical) industry may be more (less) exposed to significant cyber losses. Loss events after 2014 have been more prominent than before, consistent with the fact that claims have increased and profitability has decreased in the current cyber-insurance market (A.M. Best, 2019).

Concerning calculating the cost of claims, the financial impact of the size of the data breach is statistically significant across quantiles from 5% to 95%, which implies that the size of a breach is a key determinant of the final claims. We find higher costs per record for smaller loss events (i.e., low loss distribution quantiles) and lower costs per record for large loss events. This indicates that small firms might have higher costs, leading to higher costs per record for a breach event. We also show that cyber losses are significantly affected by legal risks. Thus, insurers would need to regard cyber insurance as a potential long-tail line and consider the legal cost for the premium calculation.

Our estimates at the 5% and 95% quantiles deviate significantly from the estimate of the OLS model. This finding contributes to studies that mostly use the OLS estimation for the claims cost of cyber risk events and demonstrates the importance of considering heterogeneous impacts. It also implies that the claims calculators of cyber insurers based on their simplified models may over- or under-estimate the final claims, which can cause insurers and insured individuals to suffer from financial mismanagement. In this sense, our study can add value to the practical effort to identify an optimal method that can be used to estimate appropriate cyber claims.

We consider the heterogeneous impacts of firm-specific factors on insurance pricing by comparing pure premiums from the quantile regression estimation (i.e., QPP) with those from a two-part GLM and the Tweedie model. We identify heterogeneous effects on insurance premiums by considering the heterogeneous probabilities of claims for firms with different sizes and relative security levels in different industries. We observe that, generally, quantile-based premiums are larger than those from the two-part GLM and the Tweedie model, particularly for large firms. We also show that firms more likely to have a higher probability of claims (relatively higher exposure) and lower security than others need to pay higher premiums. This implies that such quantile-based pricing models could help cyber insurers differentiate risks by taking different quantiles with heterogeneous firm sizes, security levels, and claims probabilities.

As the cyber-insurance market is still growing rapidly, understanding cyber events and their financial impacts should be improved significantly. A more expanded database is needed to achieve this goal. A large database may have more firm-specific factors valid for insurance pricing and determination of final claims. More detailed information on individual cyber security at the firm level might yield useful insights into the interaction between prevention and losses (i.e., self-protection vs. risk transfer). Our study provides the first analysis in this direction. However, more detailed security information is needed to verify and better understand the impact of cyber security on loss frequency and severity. Additionally, cyber risk has changed rapidly over time. As it continues to change, the risk landscape that insurers and their clients must deal with changes accordingly as the business environment and societies become hyper-connected. This change would make it more challenging for cyber insurers to develop a sophisticated predictive model because predictors also change with the development of information technology. The continuous evolution of the cyber risk landscape thus remains a fruitful and important area for future research in both the insurance mathematics and economics domains.

Declaration of competing interest

None declared.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT: Ministry of Science and ICT) (No. 2021R1G1A1008032).

References

A.M. Best, 2019. Cyber insurers are profitable today, but wary of tomorrow's risks. A.M. Best, New Jersey.

AIR Worldwide, 2017. Insuring Cyber Risk. AIR Worldwide, Boston.

Aiyer, B., Anant, V., Di Mattia, D., 2021. Securing Small and Medium-Size Enterprises: What's Next? McKinsey & Company. Retrieved from https://www.mckinsey.com/~/media/mckinsey/business%20functions/risk/our%20insights/securing%20small%20and%20medium%20size%20enterprises%20whats%20next/securing-small-and-medium-size-enterprises-whats-next-vf.pdf.

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

Baione, F., Biancalana, D., 2019. An individual risk model for premium calculation based on quantile: a comparison between generalized linear models and quantile regression. North American Actuarial Journal 23 (4), 573–590.

Biener, C., Eling, M., Wirfs, J.H., 2015. Insurability of cyber risk: an empirical analysis. The Geneva Papers on Risk and Insurance. Issues and Practice 40 (1), 131–158.

Department for Digital, Culture, Media & Sport (DCMS), 2021. Cyber Security Breaches Survey 2018. U.K. Government, London.

Dreyer, P., Jones, T., Klima, K., Oberholtzer, J., Strong, A., Welburn, J.W., Winkelman, Z., 2018. Estimating the Global Cost of Cyber Risk. RAND Corporation, Santa Monica.

Edwards, B., Hofmeyr, S., Forrest, S., 2016. Hype and heavy tails: a closer look at data breaches. Journal of Cybersecurity 2 (1), 3–14.

EIOPA, 2018. Understanding Cyber Insurance – A Structured Dialogue with Insurance Companies. European Insurance and Occupational Pensions Authority, Luxembourg.

Eling, M., Jung, K., 2018. Copula approaches for modeling cross-sectional dependence of data breach losses. Insurance. Mathematics & Economics 82, 167–180.

Eling, M., Loperfido, N., 2017. Data breaches: goodness of fit, pricing, and risk measurement. Insurance. Mathematics & Economics 75, 126–136.

Eling, M., Schnell, W., 2016. What do we know about cyber risk and cyber risk insurance? The Journal of Risk Finance 17 (5), 474–491.

Eling, M., Schnell, W., 2020. Capital requirements for cyber risk and cyber risk insurance: an analysis of solvency II, the US Risk-based capital standards, and the swiss solvency test. North American Actuarial Journal 24 (3), 370–392.

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

Franke, U., 2017. The cyber insurance market in Sweden. Computers & Security 68, 130–144.

Frees, E., 2009. Regression Modeling with Actuarial and Financial Applications. Cambridge University Press, New York.

Fung, B., 2017. Actually, every single Yahoo account got hacked in 2013. Retrieved from The Washington Post. October 4. https://www.washingtonpost.com/news/the-switch/wp/2017/10/03/yahoos-2013-data-breach-affected-all-3-billion-accounts-tripling-its-previous-estimate/.

Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance. Mathematics & Economics 70, 205–215.

Gordon, L.A., Smith, R., 2007. Incentives for Improving Cybersecurity in the Private Sector: A Cost-Benefit Perspective. Congressional Testimony.

Gordon, L.A., Loeb, M.P., Lucyshyn, W., Zhou, L., 2015. Increasing cybersecurity investments in private sector firms. J. Cybersecur. 1 (1), 3–17.

Gordon, L.A., Loeb, M.P., Lucyshyn, W., Zhou, L., 2018. Empirical evidence on the determinants of cybersecurity investments in private sector firms. Journal of Information Security 9 (2), 133–153.

Haberman, S., Renshaw, A.E., 1996. Generalized linear models and actuarial science. Journal of the Royal Statistical Society. Series D. The Statistician 45 (4), 407–436.

Heras, A., Moreno, I., Vilar-Zanón, J., 2018. An application of two-stage quantile regression to insurance ratemaking. Scandinavian Actuarial Journal 2018 (9), 753–769.

Hiscox, 2019. Hiscox cyber readiness report 2019. Hiscox, Bermuda.

Hsiao, C., Kim, C., Taylor, G., 1990. A statistical perspective on insurance rate-making. Journal of Econometrics 44 (1–2), 5–24.
Jacobs, J., 2014. Analyzing ponemon cost of data breach. Retrieved from Data Driven Security. https://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/. December 11.
Jørgensen, B., de Souza, M., 1994. Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal 1994 (1), 69–93.
Jung, K., 2021. Extreme data breach losses: an alternative approach to estimating probable maximum loss for data breach risk. North American Actuarial Journal 25 (4), 580–603.
Keasey, K., Short, H., 1990. The accounting burdens facing small firms: an empirical research note. Accounting and Business Research 20 (80), 307–313.
Kocherginsky, M., He, X., Mu, Y., 2005. Practical confidence intervals for regression quantiles. Journal of Computational and Graphical Statistics 14 (1), 41–55.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.
Koenker, R., Hallock, K., 2001. Quantile regression. The Journal of Economic Perspectives 15 (4), 143–156.
Kudryavtsev, A., 2009. Using quantile regression for rate-making. Insurance. Mathematics & Economics 45 (2), 296–304.
Lee, A., 2013. Welcome to the Unicorn Club: Learning from Billion-Dollar Startups. TechCrunch. November 3. Retrieved from https://techcrunch.com/2013/11/02/welcome-to-the-unicorn-club/.
Leong, Y.-Y., Chen, Y.-C., 2020. Cyber risk cost and management in IoT devices-linked health insurance. The Geneva Papers on Risk and Insurance. Issues and Practice 45, 737–759.
Lloyd's, 2017. Counting the cost: Cyber exposure decoded. Lloyd's in cooperation with Cyence, London.
Maillart, T., Sornette, D., 2010. Heavy-tailed distribution of cyber-risks. The European Physical Journal. B, Condensed Matter Physics 75 (3), 357–364.
McCoy, K., 2017. Target to pay $18.5M for 2013 data breach that affected 41 million consumers. USA today. Retrieved from https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/. May 23.
McLean, R., 2019. A hacker gained access to 100 million Capital One credit card applications and accounts. CNN. Retrieved from https://edition.cnn.com/2019/07/29/business/capital-one-data-breach/index.html. July 30.
Muermann, A., Kunreuther, H., 2008. Self-protection and insurance with interdependencies. Journal of Risk and Uncertainty 36 (2), 103–123.
Munich Re, 2020. Cyber insurance: risks and trends 2020. Retrieved from Munich Re https://www.munichre.com/topics-online/en/digitalisation/cyber/cyber-insurance-risks-and-trends-2020.html. April 14.
OECD, 2017. Enhancing the Role of Insurance in Cyber Risk Management. OECD Publishing, Paris.
Ohlsson, E., Johansson, B., 2010. Non-life Insurance Pricing with Generalized Linear Models. Springer, Berlin.
Palsson, K., Gudmundsson, S., Shetty, S., 2020. Analysis of the impact of cyber events for cyber insurance. The Geneva Papers on Risk and Insurance. Issues and Practice 45, 564–579.
Peng, C., Xu, M., Xu, S., Hu, T., 2017. Modeling and predicting extreme cyber attack rates via marked point processes. Journal of Applied Statistics 44 (14), 2534–2563.
Ponemon Institute, 2019a. Cost of a data breach report 2019. Ponemon Institute, Michigan.
Ponemon Institute, 2019b. Global State of Cybersecurity in Small and Medium-Sized Businesses. Ponemon Institute, Michigan.
Ponemon Institute, 2020. Cost of a data breach report 2020. Ponemon Institute, Michigan.
PriceWaterhouseCoopers (PwC), 2019. Insurance Banana Skins 2019. PwC, London.
Revzin, E., Majumdara, D., Bassett Jr, G.W., 2014. Conditional quantile regression models of melanoma tumor growth curves for assessing treatment effect in small sample studies. Statistics in Medicine 33 (29), 5209–5220.
Romanosky, S., 2016. Examining the costs and causes of cyber incidents. J. Cybersecur. 2 (2), 121–135.
Romanosky, S., Ablon, L., Kuehn, A., Jones, T., 2019. Content analysis of cyber insurance policies: how do carriers price cyber risk? J. Cybersecur. 5 (1), 1–19.
Shi, P., 2016. Insurance ratemaking using a copula-based multivariate Tweedie model. Scandinavian Actuarial Journal 2016 (3), 198–215.
Steinberg, S., 2019. Cyberattacks now cost companies $200,000 on average, putting many out of business. CNBC. Retrieved from https://www.cnbc.com/2019/10/13/cyberattacks-cost-small-companies-200k-putting-many-out-of-business.html. October 13.
Tarr, G., 2012. Small sample performance of quantile regression confidence intervals. Journal of Statistical Computation and Simulation 82 (1), 81–94.
The Federal Trade Commission, 2020. Equifax data breach settlement. Retrieved from the Federal Trade Commission. https://www.ftc.gov/enforcement/cases-proceedings/refunds/equifax-data-breach-settlement. January.
Tidy, J., 2020. Marriott Hotels fined £18.4m for data breach that hit millions. BBC. October 30. Retrieved from. https://www.bbc.com/news/technology-54748843#:~:text=The%20UK's%20data%20privacy%20watchdog,compromised%20in%20a%20cyber%2Dattack.
Tweedie, M., 1984. An index which distinguishes between some important exponential families. In: Statistics: Applications and New Directions: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Vol. 579. Indian Statistical Institute, Calcutta, pp. 579–604.
Wheatley, S., Maillart, T., Sornette, D., 2016. The extreme risk of personal data breaches and the erosion of privacy. The European Physical Journal. B, Condensed Matter Physics 89, 7.
Xie, X., Lee, C., Eling, M., 2020. Cyber insurance offering and performance: an analysis of the US cyber insurance market. The Geneva Papers on Risk and Insurance. Issues and Practice 45 (4), 690–736.
Zeller, G., Scherer, M., 2021. A comprehensive model for cyber risk based on marked point processes and its application to insurance. European Actuarial Journal, 1–53.
Zhao, X., Xue, L., Whinston, A.B., 2013. Managing interdependent information security risks: cyberinsurance, managed security services, and risk pooling arrangements. Journal of Management Information Systems 30 (1), 123–152.