

Wang, Kaiwen et al.

Article

Treatment level and store level analyses of healthcare data

Risks

Provided in Cooperation with:

MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Wang, Kaiwen et al. (2019) : Treatment level and store level analyses of healthcare data, *Risks*, ISSN 2227-9091, MDPI, Basel, Vol. 7, Iss. 2, pp. 1-22, <https://doi.org/10.3390/risks7020043>

This Version is available at:

<https://hdl.handle.net/10419/257881>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.


You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>

Treatment Level and Store Level Analyses of Healthcare Data

Kaiwen Wang ^{1,2}, Jiehui Ding ^{1,2}, Kristen R. Lidwell ¹, Scott Manski ¹, Gee Y. Lee ^{1,2,*} and Emilio Xavier Esposito ³ 

¹ Department of Statistics and Probability, Michigan State University, C413 Wells Hall, 619 Red Cedar Rd, East Lansing, MI 48824, USA; wangka15@msu.edu (K.W.); dingjeh@msu.edu (J.D.); lidwellk@msu.edu (K.R.L.); manskisc@stt.msu.edu (S.M.)

² Department of Mathematics, Michigan State University, C212 Wells Hall, 619 Red Cedar Rd, East Lansing, MI 48824, USA

³ exeResearch, LLC, 32 University Dr, East Lansing, MI 48823, USA; emilio@exeResearch.com

* Correspondence: leegee@msu.edu; Tel.: +1-517-353-6332

Received: 23 February 2019; Accepted: 13 April 2019; Published: 17 April 2019



Abstract: The presented research discusses general approaches to analyze and model healthcare data at the treatment level and at the store level. The paper consists of two parts: (1) a general analysis method for store-level product sales of an organization and (2) a treatment-level analysis method of healthcare expenditures. In the first part, our goal is to develop a modeling framework to help understand the factors influencing the sales volume of stores maintained by a healthcare organization. In the second part of the paper, we demonstrate a treatment-level approach to modeling healthcare expenditures. In this part, we aim to improve the operational-level management of a healthcare provider by predicting the total cost of medical services. From this perspective, treatment-level analyses of medical expenditures may help provide a micro-level approach to predicting the total amount of expenditures for a healthcare provider. We present a model for analyzing a specific type of medical data, which may arise commonly in a healthcare provider's standardized database. We do this by using an extension of the frequency-severity approach to modeling insurance expenditures from the actuarial science literature.

Keywords: medical data analysis; store sales analysis; predictive modeling; generalized additive models

1. Introduction

Frequency-severity models are foundational to insurance claims' modeling for actuarial applications. In a frequency-severity approach, the analyst would utilize a model for the frequency of insurance claims, as well as a model for the severity of insurance claims. By doing this, the modeler is able to understand the factors that influence the frequency and the severity of insurance claims. This paper illustrates how the frequency-severity modeling approach can be extended, so that it is capable of modeling healthcare expenditures at the treatment level. The idea is to not only consider the frequency of the patients, but also the frequency of the treatments incurred by each patient at each department of a healthcare network. This allows the analyst to understand the variation of the frequency of patients, the frequency of treatments at each department, as well as the severity of the expenditures, via the coefficients estimated from the analysis.

In order to illustrate our approach, we first motivate the frequency-severity modeling approach from basic principles. Then, we explain how the framework can be extended to a treatment-level analysis. In this process, the generalized additive models (GAM) approach can be helpful for capturing non-linear relationships with the response variable and the explanatory variables, and we demonstrate

how this can be done. We then utilize a synthetically-generated dataset to illustrate our approach. The synthetic data have been designed to mimic the treatment-level dataset of a generic healthcare provider with multiple departments in a network of clinics. Our approach can be helpful for analyzing datasets of a similar nature. In particular, we found it useful for analyzing the store-level data and treatment-level data obtained from a local non-profit healthcare organization.

The paper proceeds in the following order: In Section 2, a review of relevant literature is provided. Section 3 provides an overview of lognormal regression and how it can be applied to a store-level data. Section 4 introduces the reader to GLM models and the motivation for their use on the store data at the county level. In our project, the discussed county-level data were obtained by merging the store data and the total annual payroll data at the county level using data from the U.S. Census Bureau. Section 5.1 describes in detail a treatment-level model that can be applied to generic medical treatment data. In Section 5.2, an overview of how the treatments can be categorized is provided. Section 6 explains how the treatment level model can be extended using a GAM framework in order to capture the nonlinear relationship between the time variable and the response variable. Section 7 summarizes our finding for the treatment level analysis, and Section 7.3 describes model validation approaches. Section 8 concludes the paper with final remarks and possible future work.

2. Literature

2.1. Frequency-Severity Modeling

The regression framework in this paper follows closely the frequency-severity approach in the actuarial science literature, where an overview of dependent frequency-severity modeling is provided by [Frees et al. \(2016\)](#). For an overview of regression modeling in the actuarial science context, the reader is referred to the [Frees \(2009\)](#) monograph *Modeling with Actuarial and Financial Applications*. According to Frees and coworkers [Frees \(2014\)](#); [Frees et al. \(2016\)](#), there are good reasons for modeling the frequency and the severity (cost) of claims separately. Insurers may impose coverage modifications, such as the application of deductibles, co-insurance, or policy limits on a per occurrence basis. Only knowing each policy's aggregate claim amount does not allow for the calculation of mean expenditures under per occurrence-based coverage modifications. Meanwhile, covariates that aid the understanding of the outcomes can differ substantially between the frequency and the severity. In actuarial applications, different risk mitigation strategies may be suggested depending on the coefficient obtained for the frequency part and the severity part. In addition, the data files encountered in practice (analysis and modeling) often encourage developing independent frequency and severity models. Furthermore, in actuarial applications, it is important to consider the fact that regulators may typically require the inclusion of the number of claims along with the amounts (costs) within reports. Finally, the use of a separate severity model allows complex features for the marginal models to be captured, such as the heavy-tailed nature of the severity distribution.

The recent trend in actuarial science research is to incorporate sophisticated marginal distributions into the frequency-severity framework, allowing more flexibility in the modeling. For an overview of heavy-tail data modeling, the reader is referred to works by [Sun et al. \(2008\)](#), [Yang \(2011\)](#), and [Shi \(2014\)](#). Focusing on the treatment frequency aspect, zero-inflated models have been utilized in insurance modeling, and the reader is directed to the excellent overview by [Boucher \(2014\)](#). The zero-inflated model has been extended to zero-one-inflated models by [Frees et al. \(2016\)](#). Such sophisticated marginal models allow for specific features of the data to be captured within the models.

2.2. Longitudinal Modeling

The response and explanatory variables observed in a medical treatment-level analysis may be longitudinal. This means observations occur and recur over time with a certain correlation structure among the observations. The reader is referred to Frees' book—*Longitudinal and Panel Data: Analysis and Applications in the Social Sciences* [Frees \(2004\)](#) for a primer on longitudinal data analysis. Recently,

copula models have been used to model longitudinal data, and [Ruscone and Osmetti \(2016\)](#), [Shi and Valdez \(2014\)](#), and [Shi \(2012\)](#) have separately demonstrated copula's ability to model longitudinal data. The application of pair-copulas can help model complex dependence structures as in the work of [Smith et al. \(2010\)](#).

The hierarchical modeling framework utilized in the treatment level analysis of this article is related to the work of [Frees and Valdez \(2008\)](#). They demonstrated the ability to model insurance claims using a stochastic variable to represent the yearly number of claims, another stochastic variable to denote the claim type, and a third stochastic variable to stand for the claim amount. These variables are observed for each observational unit $\{it\}$, where i corresponds to risk class and t represents calendar year. In our research, time was represented by days, weeks, or months, depending on how the data were tabulated. Different tabulations may result in interesting discoveries regarding the clinics.

2.3. Medical Data Analysis

A detailed example of applying regression techniques on to healthcare expenditure modeling is demonstrated in the work of [Frees et al. \(2011\)](#). In the health economics literature, [Keeler and Rolph \(1988\)](#) studied insurance claims from the RANDHealth Insurance Experiment and approached the data analysis in a similar way as we present herein. Keeler and Rolph's goal was to analyze the effects of health insurance plans on expenditures using a random coefficients model along with logarithmic expenditures and count distributions to model the episode frequencies. In a related paper by [Rosenberg and Farrell \(2008\)](#), a Bayesian approach was applied to model the inpatient utilization and expenditures at the individual level. The primary goal of these noted studies was the prediction of expenditures (costs). Our work also focuses on the prediction problem, yet we are also interested in the coefficient estimates of certain explanatory variables to understand their relationship with the response variables.

3. Lognormal Regression

3.1. Model

Imagine observing a certain response variable, which is defined on the domain $(0, \infty)$. Examples of such responses may be insurance loss amounts, medical expenditure amounts, or property and crop losses due to natural disasters. One way to model this type of response is to use a lognormal variable, because the lognormal distribution is defined over $(0, \infty)$. In this case, linear modeling is a powerful tool for understanding the relationship between a response variable $\ln Y$ and the explanatory (independent) variables X_1, X_2, \dots, X_p . If observations $\ln y_1, \dots, \ln y_n$ are obtained from random realizations of the variable $\ln Y$, along with the corresponding explanatory variables $x_{11}, x_{21}, \dots, x_{n1}$ for X_1 , and $x_{12}, x_{22}, \dots, x_{n2}$ for X_2 , all the way up to $x_{1p}, x_{2p}, \dots, x_{np}$ for X_p , then we use matrix algebra to express the relationship between the responses and the explanatory variables by:

$$\ln \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or in other words:

$$\ln y = X\beta + \epsilon$$

Different assumptions on ϵ result in different models. The standard linear regression model assumes each component of ϵ is normally distributed or, equivalently, each component of $\ln y$ is normally distributed. In this case, assuming the components of $\ln y$ are independent, we obtain:

$$\ln y_i \sim N(x_i^T \beta, \sigma^2),$$

where $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $\boldsymbol{\beta}$ and σ are parameters to be estimated. Once normality is assumed, linear regression can be performed to estimate the parameters $\boldsymbol{\beta}$ and σ . The parameters may be estimated using various approaches, but the maximum likelihood estimation (MLE) is considered a standard approach. Intuitively, the probabilities for observing each data point are:

$$f(y_i, \mathbf{x}_i) = \frac{1}{y_i \sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (\ln y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right],$$

and multiplied out to form:

$$L = \prod_{i=1}^n f(y_i, \mathbf{x}_i) = \frac{1}{y_1 \dots y_n (2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\ln \mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\ln \mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right].$$

The logarithm of the expression above is called the log-likelihood. The log-likelihood $\ln L$ may be maximized by setting its derivative with respect to the parameter $\boldsymbol{\beta}$ to zero.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln L = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{\beta}} (\ln \mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\ln \mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

Solving the above system analytically is a standard exercise in vector calculus, and derivations are found in Generalized Linear Models with Applications in Engineering and the Sciences, by Myers et al. (2002). The result is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\ln \mathbf{y}),$$

which is a familiar expression for those who have a background in linear regression. The important point is that standard linear regression can be understood as a procedure, where some log-likelihood function is maximized. Numerically minimizing the function $-\ln L$, with respect to the parameter $\boldsymbol{\beta}$, results in a nearly identical answer to those found analytically. Notice that the mean of a lognormally-distributed random variable with parameters μ_i and σ is:

$$E[y_i] = \exp \left(\mu_i + \frac{\sigma^2}{2} \right).$$

Hence, for predictive applications, we have:

$$\hat{y}_i = \exp \left(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) \cdot \exp \left(\frac{\hat{\sigma}^2}{2} \right), \quad (1)$$

which takes on values within $(0, \infty)$. Notice that the log transform of the response variable assures the resulting predictions to be positive values.

3.2. Specific Example

In this section, we present a model that can be used for the analysis of hypothetical store-level data of a healthcare provider. The total square footage of the stores and the population within certain miles of the stores are used as independent variables for the regression modeling. In this case, our model specification is:

$$E[y_i] = \exp (\beta_0 + \beta_1 \ln x_{i1} + \beta_2 \ln x_{i2}) \cdot \exp \left(\frac{\sigma^2}{2} \right),$$

where y_i is store sales, x_{i1} is the total square feet of the i th observation, and x_{i2} is the population within certain miles of the i th observation. Collinearity may arise if the two variables x_{i1} and x_{i2} are highly correlated. In our analysis, due to the collinearity of these two variables, only one of the variables was significant at the 0.05 level (95% confidence). The adjusted R^2 of the model can be used

to determine the fit of the model, where a high R^2 may indicate that a large portion of the variability in the response variable could be explained by the explanatory variables. The reason why the log amounts of the square feet and population are used instead of the provided observed values is because the coefficients have an elasticity interpretation when included as log amounts. We see that:

$$\frac{\partial E[y_i]}{\partial x_{i,1}} = \exp(\beta_0 + \beta_1 \ln x_{i,1} + \beta_2 \ln x_{i,2}) \cdot \exp\left(\frac{\sigma^2}{2}\right) \cdot \frac{\beta_1}{x_{i,1}} = \frac{E[y_i]}{x_{i,1}} \beta_1$$

Solving this for β_1 , we have:

$$\beta_1 = \frac{\partial E[y_i]/E[y_i]}{\partial x_{i,1}/x_{i,1}} = \frac{\% \Delta E[y_i]}{\% \Delta x_{i,1}} = \frac{\text{percent change in } E[y_i]}{\text{percent change in } x_{i,1}}$$

and hence, β_1 is the total square feet elasticity of the expected sales. A similar interpretation is possible for β_2 , and hence, we use the log amounts of the explanatory variables along with a log-link. The analyst may utilize Q-Q (quantile-quantile) plots to assess the goodness of the fit of the model. A Q-Q plot is a graphical approach to comparing two probability distributions by plotting their quantiles against each other. During our project, we also tried fitting the gamma model (explained later in the paper), and the Q-Q plot was not significantly different from the lognormal model for our particular dataset.

4. GLMs

4.1. Motivation

The model shown in Section 3.2 helps us tell an interesting story; however, it does not capture the correlation between annual payroll and the location of the stores. For this, we perform an analysis at the county level, meaning that the unit of analysis is a single lattice and the response variable is the number of stores observed in each area. For each lattice, we count the number of stores within the area and create a response variable using the counts. Along with the response variable, the total annual payroll for each county can be obtained by aggregating the payroll data obtained (downloaded) from the U.S. Census Bureau.

The problem with the lognormal approach to modeling the store counts at the county level is that the log response may no longer be assumed to have a normal distribution. Because the response is the number of stores in a county, it is no longer a real number in the range $(0, \infty)$, but instead numbers in $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$. We used generalized linear models (GLMs), where the regression technique is “generalized” to response variables that are non-normal. The technique was first introduced by [Nelder and Wedderburn \(1972\)](#), and today, there are numerous books covering the topic; the reader is directed to the works of [Myers et al. \(2002\)](#), [Ohlsson and Johansson \(2010\)](#), and [Dobson \(2008\)](#). Because of GLM’s inherent flexibility, they have been the workhorse model within the insurance industry where modeling non-normal responses such as insurance claim counts, claim indicators, or claim amounts is a major part of the actuarial analyst’s workflow.

Today, predictive analytics modeling has become a core component of the daily tasks performed by actuarial analysts. With the abundance of data and the availability of standardized statistical routines, actuaries are now able to analyze insurance claims’ data in a systematic way. The GLM framework has become a central component among the tools used by actuaries to analyze insurance claims’ data. Some advantages of GLM are as follows:

- GLMs are able to model both the claim frequency and the claim severity in a unified language similar to that used in linear regression modeling.
- Standardized routines are readily available for software packages such as R, SAS, SPSS, and JMP, allowing the analyst to avoid writing complex maximum likelihood code and scripts.
- GLMs have flexibility in incorporating covariates into the modeling framework.

One of the main goals of predictive analytics modeling, in an actuarial context, is to provide a rating engine for an insurance provider. A rating engine is a way to express the relationship between explanatory variables and the response variable of interest. The coefficient estimates obtained from regression modeling can be used in the rating engine for this purpose. We believe the GLM framework is also useful for medical data analysis, as we will demonstrate.

4.2. Poisson Regression for Count Data

A GLM model can be understood as the application of the exponential family distributions to a regression problem. An exponential family distribution has the form:

$$f(y) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (2)$$

Here, θ is the location parameter and ϕ is the dispersion parameter. To obtain the normal distribution from this, set $\theta = \mu$, $b(\theta) = \mu^2/2$, $a(\phi) = \phi = \sigma^2$, and:

$$c(y) = -\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$

Hence, the normal distribution is an exponential family distribution. Another member of the exponential family is the Poisson distribution that can be obtained by setting $\theta = \ln \mu$, $b(\theta) = e^\theta$, and $c(y) = -\ln(y!)$. For the Poisson case, the probability function reduces to:

$$f(y) = \exp [y \ln \mu - \mu - \ln(y!)] = \frac{e^{-\mu} \mu^y}{y!}$$

The probability function of the Poisson model gives the probability of an observation within \mathbb{Z}^+ . Hence, the Poisson distribution is used to model counts of events. Now, we informally define the GLM model. Assume the observations y_1, y_2, \dots, y_n are independent response variables following the exponential family distribution, with means $\mu_1, \mu_2, \dots, \mu_n$, along with explanatory variables $x_{11}, x_{21}, \dots, x_{n1}$ for the first explanatory variable, $x_{12}, x_{22}, \dots, x_{n2}$ for the second explanatory variable, and so on. We use a link function to parameterize the mean of the response variables, so that:

$$E[y_i] = \mu_i = g^{-1} \left(x_i^T \beta \right)$$

where g is a monotonic differentiable function. We use the log-link for the Poisson model, so that $g(\cdot) = \log(\cdot)$. Once the mean of the distribution is parameterized, we apply the maximum likelihood estimation to obtain an estimate for the parameter β . Specifically, the log-likelihood:

$$\ln L = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

is minimized numerically. The important point here is that by parameterizing the mean of the response, maximizing the log-likelihood defined by the exponential family distribution turns into essentially a regression problem. Hence, using this approach, we were able to regress the count variable (number of stores in a county) onto a set of explanatory variables x_i .

4.3. Other GLM Models

In Equation (2), setting $\theta = -1/\mu$, $a(\phi) = r^{-1}$, $b(\theta) = -\ln(-\theta)$, and $c(\phi) = r \ln r - \ln \Gamma(r) + (r-1) \ln y$, where r is a shape parameter satisfying $\mu = r\lambda$ for a scale parameter λ , we have the gamma distribution:

$$f(y) = \frac{1}{\Gamma(r)} \left(\frac{1}{\lambda} \right)^r e^{-y/\lambda} y^{r-1}$$

The gamma distribution is often used with a log-link, so that:

$$E[y_i] = \mu_i = \exp(x_i^T \beta)$$

is the parameterization of the mean, resulting in a regression framework. The support of the gamma distribution is $(0, \infty)$, and hence, the gamma distribution may be used to replace the log-normal model to model insurance claim amounts, treatment amounts, and expenditure sizes. The advantage of the gamma distribution is its theoretical properties when it comes to the sum of independent identically-distributed gamma random variables. The Poisson sum of gamma random variables is called the Tweedie distribution in the literature; see [Ohlsson and Johansson \(2010\)](#). [Frees and Lee \(2016\)](#) have demonstrated the ability to model endorsement premiums via a Tweedie distribution.

Meanwhile, setting $\theta = \ln(\pi/(1-\pi))$, $a(\phi) = 1$, $b(\theta) = n \ln(1 + e^\theta)$, $c(\phi) = \ln \binom{n}{y}$ results in the binomial distribution. The probability function for the binomial distribution with $n = 1$ is:

$$f(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} = \begin{cases} \pi & \text{for } y = 1 \\ 1 - \pi & \text{for } y = 0 \end{cases}$$

The support of the binomial distribution is $\{0, 1\}$, meaning that each observation is either true or false. In a predictive application, one is often interested in calculating the probability of a true response; hence, we parameterize π using a logit link function.

$$x_i^T \beta = \text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \implies \pi_i = \text{logit}^{-1}(x_i^T \beta) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

Again, notice that we have formulated a regression framework, for the case where the response variable y_i takes on values within $\{0, 1\}$. Now, we have a number of tools in our toolbox: we are able to regress binary response variables and count response variables and continuous response variables onto a set of explanatory variables, using a unified regression framework. Next, let us understand the building blocks of a hierarchical model. The most basic hierarchical model is a frequency-severity modeling framework, as explained in [Section 4.4](#).

4.4. The Frequency-Severity Model

In the actuarial literature, it is common for an insurance claims model to be applied to a dataset, in which the following variables are observed:

- N , the number of claims (events),
- Y_i , $i = 1, \dots, N$, the amount of each claim (expense).

In this case, the aggregate amount of the claims for the insurance company, or healthcare provider, becomes:

$$S = Y_1 + \dots + Y_N$$

and if we assume independence between the frequency (N) and the severity (Y_i) of the claims, then we can model the two components separately and have:

$$E[S] = E[N] \cdot E[Y_i]$$

According to [Frees et al. \(2016\)](#), the aggregate claim amount S is an important feature for an insurer's balance sheet because it is the amount paid on claims. The same may be true for a non-profit organization, which might interpret the healthcare expenditures of individuals as "claims", and the total amount of such expenditures as the cost of operating for the healthcare provider. While the aggregate amount S is often of interest, for an actuarial application, there are benefits to modeling

the frequency and severity of the losses separately. If the losses represent healthcare expenditures paid by a healthcare provider, then we may imagine the model as representing the following situation: an individual's decision to use healthcare may vary (the frequency), and the individual's cost (the severity) is likely related to the characteristics of the physician or the type of treatment (healthcare provider). Herein, we take the position that the joint modeling of frequency and severity of claims is the point of interest. The long history of studying frequency, severity, and the aggregate claim for independent and identically-distributed realizations of random variables is a cornerstone of actuarial science. For an introduction to actuarial science, we direct the reader to the introductory textbook—Loss Models: From Data to Decisions by [Klugman et al. \(2012\)](#).

We assume the (actuarial) analyst has access to the explanatory variables. For an analyst in the insurance industry, the independent variables—various characteristics of the policyholder—are obtained from the policyholder's application form. Explanatory variables for auto insurance may include the driver's age, vehicle type, and region of operation. Recently, there have been attempts to incorporate telematics data into insurance rate-making, where distance driven, distance driven in the city (city mileage), and how often the vehicle's operator violates speed limits are incorporated into the modeling of the claim frequencies and severities. Additionally, it has been demonstrated that a person's credit score is an important predictor of auto claims; see Frees' review [Frees \(2015\)](#) for further discussion.

In healthcare applications, the independent variables may be obtained from the patient's clinic visitation history. For example, the age and gender of the patient may be likely predictors for the frequency of certain medical treatments. Meanwhile, independent of the patient, the treatment types and frequencies may also be influenced by the season (month of the year) or depend on which day of the week administered. The severity of the healthcare expenditures is related to the International Classification of Diseases (ICD) code chapter, sub-chapter, and major designation of the treatment. Unfortunately, such patient-level explanatory variables are often difficult to obtain for academic research because they are the private information of the patients and protected by the Health Insurance Portability and Accountability Act (HIPAA). Based on these restraints, we present a study of a more complex hierarchical model for a treatment-level analysis of healthcare expenditures for general analyses.

5. Treatment-Level Analysis

5.1. Model

In order to present a generic treatment-level model, we follow a hierarchical model, where the total cost is determined by components. The first is the number of patients arriving at a given time, and we call this random variable N_{tk} , where t is discretized time in days and k is the treatment category. We assume that treatments are provided through three hypothetical departments within the healthcare provider network. The departments may be categorized into any different number of categories to suit the particular data in hand. Given a realization of N_{tk} , each patient department (Departments 1, 2, and 3) will generate the following number of treatments for each patient $i \in 1, \dots, N_{tk}$:

- $M_{1,tki}$ number of treatments in Department 1 (the department of interest)
- $M_{2,tki}$ number of treatments in Department 2 (a department being compared with Department 1)
- $M_{3,tki}$ number of treatments in Department 3 (all other departments)

For the first category, given a realization of $M_{1,tki}$, each treatment will correspond to an ICD-10 code chapter, which may result in either a zero or positive cost. Here, j is the index of the treatment. For an arbitrary treatment in the ICD-10 code category k , for patient i at time t , the charge for the treatment can be positive or negative. Thus, we use a binary variable $P_{1,tkij}$ to model the variable. Given that the cost is positive, we let $Y_{1,tkij}$ be the random variable for the cost for treatment j of patient i at time t in treatment category k . Similar definitions are applicable for Departments 2 and 3.

Then, we define $TC_{1,tki}$ as the total cost arising from Department 1 for patient i at time t in treatment category k , and $TC_{2,tki}$, $TC_{3,tki}$ are defined similarly. In this case, we have the relationship:

$$TC_{tk} = \sum_{i=1}^{N_{tk}} [TC_{1,tki} + TC_{2,tki} + TC_{3,tki}]$$

where TC_{tk} is the total cost for the medical care provider in year t in category k . We assume that there are K different categories of treatments, as shown in Table 1 (these categories are explained in more detail in Section 5.2). We then have:

$$TC_{1,tki} = \sum_{j=1}^{M_{1,tki}} Y_{1,tkij} \quad TC_{2,tki} = \sum_{j=1}^{M_{2,tki}} Y_{2,tkij} \quad TC_{3,tki} = \sum_{j=1}^{M_{3,tki}} Y_{3,tkij}$$

Then, we have:

$$\begin{aligned} E[Y_{1,tkij}] &= E[P_{1,tkij}] \cdot E[Y_{1,tkij} | P_{1,tkij} = 1] \\ E[Y_{2,tkij}] &= E[P_{2,tkij}] \cdot E[Y_{2,tkij} | P_{2,tkij} = 1] \\ E[Y_{3,tkij}] &= E[P_{3,tkij}] \cdot E[Y_{3,tkij} | P_{3,tkij} = 1] \end{aligned}$$

where:

$$P_{1,tkij} = \begin{cases} 1 & \text{if } Y_{1,tkij} > 0 \\ 0 & \text{otherwise} \end{cases} \quad P_{2,tkij} = \begin{cases} 1 & \text{if } Y_{2,tkij} > 0 \\ 0 & \text{otherwise} \end{cases} \quad P_{3,tkij} = \begin{cases} 1 & \text{if } Y_{3,tkij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Assuming independence of N_{tk} , $M_{1,tki}$, the expected costs for each department become:

$$\begin{aligned} E[TC_{1,tki}] &= E \left[\sum_{j=1}^{M_{1,tki}} Y_{1,tkij} \right] = E[M_{1,tki}] \cdot E[Y_{1,tkij}], \\ E[TC_{2,tki}] &= E \left[\sum_{j=1}^{M_{2,tki}} Y_{2,tkij} \right] = E[M_{2,tki}] \cdot E[Y_{2,tkij}], \\ E[TC_{3,tki}] &= E \left[\sum_{j=1}^{M_{3,tki}} Y_{3,tkij} \right] = E[M_{3,tki}] \cdot E[Y_{3,tkij}], \end{aligned}$$

where we assume the independence of $M_{1,tki}$ and $Y_{1,tkij}$, and so on. Thus, the total cost for the medical provider within treatment category k becomes:

$$\begin{aligned} E[TC_{tk}] &= E \left[\sum_{i=1}^{N_{tk}} \{TC_{1,tki} + TC_{2,tki} + TC_{3,tki}\} \right] \\ &= E[N_{tk}] \cdot \left\{ E[M_{1,tki}] \cdot E[Y_{1,tkij}] + E[M_{2,tki}] \cdot E[Y_{2,tkij}] + E[M_{3,tki}] \cdot E[Y_{3,tkij}] \right\} \end{aligned}$$

5.1.1. Modeling N_{tk} Using GLMs

It is a common practice in actuarial analysis to model N_{tk} based on covariates \mathbf{x}_{tk} via generalized linear models (GLMs). A Poisson or negative binomial distribution is used for count outcomes. It is common for analysts to use zero-inflated models, as described by Boucher (2014) or De Jong and Heller (2008) to accommodate the excessive number of zeros relative to the count values greater than zero implied by these distributions. An advantage of using GLMs is the ability to express the mean in terms of the explanatory variable \mathbf{x}_{tk} where—in actuarial practice—it is common to use a logarithmic link for this function to express the mean as:

$$E[N_{tk}] = \exp(\mathbf{x}_{tk}^T \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is a vector of parameters associated with the covariates. Logarithmic link functions are typically used because they are adept at fitting the data, allow easy parameter interpretations, and fits nicely with traditional approaches used in actuarial rate-making applications such as Mildenhall's systematic relationship between minimum bias and generalized linear models [Mildenhall \(1999\)](#). Here, we assume that N_{tk} follows a Poisson distribution so that:

$$P_{N_{tk}}(n) = \frac{\lambda_{tk}^n e^{-\lambda_{tk}}}{n!}$$

A “zero-one-inflated” model may be used—as in the work of [Frees et al. \(2016\)](#)—if there are large numbers of zeros and ones in our data. The zero-one-inflated model expands on the zero-inflated method and employs two generating processes: (1) a multinomial distribution that generates structural zeros and ones and (2) a Poisson or negative binomial distribution that generates counts, some of which may be zero or one. To parameterize the probabilities for the latent variable, a logit specification may be used, and to fit the parameters, maximum likelihood estimation can be used. In our project, for simplicity of coefficient interpretation, we assumed a Poisson distribution; the covariates \mathbf{x}_{tk} (explanatory variables) are explained in Section 5.3. Because we assume the patient frequencies are independent of other random variables in the model, we can estimate the coefficient $\boldsymbol{\beta}$ separately using maximum likelihood.

5.1.2. Modeling $M_{1,tki}$, $M_{2,tki}$, $M_{3,tki}$ Using GLMs

Assuming the number of treatments is independent of the number of patients arriving in each category, we can model $M_{1,tki}$, $M_{2,tki}$, $M_{3,tki}$ using a Poisson distribution and a framework similar to that in Section 5.1.1.

$$f_{M_{1,tki}}(m) = \frac{\gamma_{1,tki}^m e^{-\gamma_{1,tki}}}{m!}, \quad f_{M_{2,tki}}(m) = \frac{\gamma_{2,tki}^m e^{-\gamma_{2,tki}}}{m!}, \quad f_{M_{3,tki}}(m) = \frac{\gamma_{3,tki}^m e^{-\gamma_{3,tki}}}{m!}$$

where:

$$\gamma_{1,tki} = \exp(\mathbf{x}_{1,tk}^T \boldsymbol{\alpha}_1), \quad \gamma_{2,tki} = \exp(\mathbf{x}_{2,tk}^T \boldsymbol{\alpha}_2), \quad \gamma_{3,tki} = \exp(\mathbf{x}_{3,tk}^T \boldsymbol{\alpha}_3)$$

where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$ are the coefficients for the number of treatments model. Notice that $\mathbf{x}_{1,tk}$, $\mathbf{x}_{2,tk}$, and $\mathbf{x}_{3,tk}$ do not have the subscript i , because our model uses explanatory variables at the tk level. Due to the independence assumption, the coefficients may be estimated using a regression routine, or maximum likelihood, separately from the other components of the model.

5.1.3. Modeling $P_{1,tkij}$, $P_{2,tkij}$, $P_{3,tkij}$ Using GLMs

Often, logit and probit forms are commonly used for binary outcomes; see [Guillén \(2014\)](#). Each treatment j in treatment category k could have zero cost, or a positive cost. We use a logistic regression model specification, where:

$$\text{logit}\{\mathbb{E}[P_{1,tkij}]\} = \mathbf{x}_{1,tk}^T \gamma_1, \quad \text{logit}\{\mathbb{E}[P_{2,tkij}]\} = \mathbf{x}_{2,tk}^T \gamma_2, \quad \text{logit}\{\mathbb{E}[P_{3,tkij}]\} = \mathbf{x}_{3,tk}^T \gamma_3$$

Here, $\gamma_1, \gamma_2, \gamma_3$ are the coefficients to be estimated. Notice that the explanatory variables $\mathbf{x}_{1,tk}$, $\mathbf{x}_{2,tk}$, and $\mathbf{x}_{3,tk}$ do not have subscripts ij . Then, for Departments 1, 2, and 3, given a specific category k , the probability of positive treatment cost becomes:

$$\pi_{1,tkij} = \frac{\exp(\mathbf{x}_{1,tk}^T \gamma_1)}{1 + \exp(\mathbf{x}_{1,tk}^T \gamma_1)}, \quad \pi_{2,tkij} = \frac{\exp(\mathbf{x}_{2,tk}^T \gamma_2)}{1 + \exp(\mathbf{x}_{2,tk}^T \gamma_2)}, \quad \pi_{3,tkij} = \frac{\exp(\mathbf{x}_{3,tk}^T \gamma_3)}{1 + \exp(\mathbf{x}_{3,tk}^T \gamma_3)}.$$

Estimating γ_1 , γ_2 , and γ_3 boils down to maximum-likelihood. In the R programming language, binary outcomes may be modeled using the binomial family within the `glm` routine. For some datasets, it may be the case that treatments always result in expenditures. In this case, the model for $P_{1,tkij}$, $P_{2,tkij}$, $P_{3,tkij}$ is not needed.

5.1.4. Modeling the Conditional Expenditure Severity

Finally, the cost of random variables can be modeled using gamma random variables. For regression purposes, the mean of the gamma random variable may be parametrized so that:

$$\begin{aligned} E[Y_{1,tkij}|P_{1,tkij} = 1] &= \exp\left(\mathbf{x}_{1,tk}^T \boldsymbol{\xi}_1\right) \\ E[Y_{2,tkij}|P_{2,tkij} = 1] &= \exp\left(\mathbf{x}_{2,tk}^T \boldsymbol{\xi}_2\right) \\ E[Y_{3,tkij}|P_{3,tkij} = 1] &= \exp\left(\mathbf{x}_{3,tk}^T \boldsymbol{\xi}_3\right) \end{aligned}$$

where $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$ are the regression coefficients and $\mathbf{x}_{1,tk}$, $\mathbf{x}_{2,tk}$, $\mathbf{x}_{3,tk}$ are explanatory variables for the k th category treatment at time t .

There are many ways to model the severity of outcomes. For a dependence model, using a latent variable that affects both frequency and loss amounts induces a positive association. Copulas are another method to model non-linear associations among random variables. A strength of the GLM approach—for insurance analysts—is that the same set of routines can be used for continuous, as well as discrete (binary) outcomes.

GLMs have become the workhorse for insurance industry analysts interested in analyzing and modeling the severity of claims. Due to the industry's primary focus on claims' severity (total cost), several alternative approaches have been explored, and [Shi \(2014\)](#) presents an excellent introduction. For example, the generalized beta of the second kind (GB2) distribution is used by [Frees et al. \(2016\)](#). The specific severity distribution to use is an empirical question, and one may employ Q-Q plots or other model diagnostic techniques to select a distribution family that fits the data best. In the research presented here, the gamma GLM approach was used for simplicity.

5.2. Treatment Categories

The ICD-10 codes were used to encode the treatment category. The ICD-10 coding structure has been produced by the World Health Organization (WHO) and is used in countries around the world. The coding structure categorizes diseases into broad categories called “Chapters”, and these broad categories are further categorized into specific disease areas termed “Sub-Chapters”. In our data, the ICD variable contained either the ICD-9 or ICD-10 code for each treatment due to the dataset spanning the United States' adoption date of 1 October 2015. The ICD variable could be standardized into ICD-10 code chapters using a custom R script. If a valid ICD-10 code is not identified, then the treatment is removed from the dataset and thus the analysis. The treatment categories (chapters) used in the analysis are provided in Table 1.

Table 1. Treatment categories (ICD-10 chapters).

Chapter	Block	Description
1	A00–B99	Certain infectious and parasitic diseases
2	C00–D48	Neoplasms
3	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4	E00–E90	Endocrine, nutritional and metabolic diseases
5	F00–F99	Mental and behavioral disorders
6	G00–G99	Diseases of the nervous system
7	H00–H59	Diseases of the eye and adnexa

Table 1. Cont.

Chapter	Block	Description
8	H60–H95	Diseases of the ear and mastoid process
9	I00–I99	Diseases of the circulatory system
10	J00–J99	Diseases of the respiratory system
11	K00–K93	Diseases of the digestive system
12	L00–L99	Diseases of the skin and subcutaneous tissue
13	M00–M99	Diseases of the musculoskeletal system and connective tissue
14	N00–N99	Diseases of the genitourinary system
15	O00–O99	Pregnancy, childbirth and the puerperium
16	P00–P96	Certain conditions originating in the perinatal period
17	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
18	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19	S00–T98	Injury, poisoning and certain other consequences of external causes
20	V01–Y98	External causes of morbidity and mortality
21	Z00–Z99	Factors influencing health status and contact with health services
22	U00–U99	Codes for special purposes

5.3. Patient Treatment Data

A set of explanatory variables (date of treatment, cost, department, and treatment category code) may be present in the dataset, and the patient-level demographic information (age, gender, ethnicity, height, and weight) may also be observed in the data. In our data, patient demographic information was not observed. Table 2 describes the explanatory variables used in the modeling presented below for the total number of patients visiting the clinics, and the number of treatments within each category of department. For the expenditure severity model, only the treatment categories are used as explanatory variables. The indicator variable `ClinicOpen` is a binary variable indicating whether Department 1 is open, and this variable has been included to study the influence of the department's operation on the number of treatments at other departments.

Table 2. Explanatory variables for the number of patients model (for N_{tk}), number of treatments model (for $M_{1,tki}$, $M_{2,tki}$, $M_{3,tki}$), and the charge amounts model (for $Y_{1,tkij}$, $Y_{2,tkij}$, $Y_{3,tkij}$).

Variable Name	Description
<code>ClinicOpen</code>	Indicator variable of whether Department 1 is open
<code>WDay</code>	A categorical variable of the weekday. (Categories: Sun, Mon, Tue, Wed, Thr, Fri, Sat)
<code>Month</code>	A categorical variable of the month. (Categories: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)
<code>Chapter</code>	A categorical variable of the treatment category. (Categories: shown in Table 1)
<code>Time</code>	Numeric variable corresponding to the current day relative to a reference time point. In our study, the reference time point is the first day in which data are available.

6. GAMs

In order to capture the nonlinear relationship between the `Time` variable and the response variables, we utilized the generalized additive models (GAM) framework, which is an extension of the GLM framework. The GLM model for the patient frequency N_{tk} is:

$$\ln \{E[N_{tk}]\} = \mathbf{x}_{tk}^T \boldsymbol{\beta}$$

In matrix form, this can be expressed as:

$$\ln \begin{pmatrix} E[N_{1,k}] \\ E[N_{2,k}] \\ \vdots \\ E[N_{T,K}] \end{pmatrix} = \begin{bmatrix} 1 & x_{1,k,1} & x_{1,k,2} & \dots & x_{1,k,p-1} & x_{1,k,p} \\ 1 & x_{2,k,1} & x_{2,k,2} & \dots & x_{2,k,p-1} & x_{2,k,p} \\ \vdots & \vdots & \ddots & & \vdots & \vdots \\ 1 & x_{T,K,1} & x_{T,K,2} & \dots & x_{T,K,p-1} & x_{T,K,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

where we assume that $x_{t,k,1} \dots x_{t,k,p-1}$ correspond to the explanatory variables other than Time and $x_{t,k,p}$ corresponds to Time. The motivation for using a GAM model is that a polynomial of $x_{t,k,p}$ may be included in the design matrix. In this case, we may include the extra terms into the model matrix using the function:

$$f(x) = x + x^2 + \dots + x^r$$

subject to the identifiability constraint

$$\sum_{k=1}^K \sum_{t=1}^T f(x_{t,k,p}) = 0$$

which states that the function f must sum to zero over the observed values of $x_{t,k,p}$; this concept is detailed on page 211 of Wood's book on GAMs [Wood \(2017\)](#). In this case, we have:

$$\ln \begin{pmatrix} E[N_{1,k}] \\ E[N_{2,k}] \\ \vdots \\ E[N_{T,K}] \end{pmatrix} = \begin{bmatrix} 1 & x_{1,k,1} & x_{1,k,2} & \dots & x_{1,k,p-1} & x_{1,k,p} & x_{1,k,p}^2 & \dots & x_{1,k,p}^r \\ 1 & x_{2,k,1} & x_{2,k,2} & \dots & x_{2,k,p-1} & x_{2,k,p} & x_{2,k,p}^2 & \dots & x_{2,k,p}^r \\ \vdots & \vdots & \ddots & & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & x_{T,K,1} & x_{T,K,2} & \dots & x_{T,K,p-1} & x_{T,K,p} & x_{T,K,p}^2 & \dots & x_{T,K,p}^r \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \\ \psi_1 \\ \vdots \\ \psi_r \end{bmatrix}$$

where r is the degree of the polynomial. In other words,

$$\ln E[y_k] = \beta_0 + \mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \mathbf{X}_{(2)}\boldsymbol{\beta}_{(2)}$$

where:

$$\mathbf{X}_{(1)} = \begin{bmatrix} 1 & x_{1,k,1} & x_{1,k,2} & \dots & x_{1,k,p-1} \\ 1 & x_{2,k,1} & x_{2,k,2} & \dots & x_{2,k,p-1} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{T,K,1} & x_{T,K,2} & \dots & x_{T,K,p-1} \end{bmatrix} \quad \mathbf{X}_{(2)} = \begin{bmatrix} x_{1,k,p} & x_{1,k,p}^2 & \dots & x_{1,k,p}^R \\ x_{2,k,p} & x_{2,k,p}^2 & \dots & x_{2,k,p}^R \\ \vdots & \vdots & \dots & \vdots \\ x_{T,K,p} & x_{T,K,p}^2 & \dots & x_{T,K,p}^R \end{bmatrix}$$

and the coefficients are:

$$\beta_0 \quad \text{and} \quad \boldsymbol{\beta}_{(1)} = (\beta_1, \beta_2, \dots, \beta_{p-1})^T \quad \text{and} \quad \boldsymbol{\beta}_{(2)} = (\psi_1, \psi_2, \dots, \psi_R)^T$$

In practice, a basis other than the polynomial may be used to implement a GAM. In this case, we first define:

$$\boldsymbol{\Phi}_{(2)} = \begin{bmatrix} B(x_{1,k,p}, 1) & B(x_{1,k,p}, 2) & \dots & B(x_{1,k,p}, R) \\ B(x_{2,k,p}, 1) & B(x_{2,k,p}, 2) & \dots & B(x_{2,k,p}, R) \\ \vdots & \vdots & \dots & \vdots \\ B(x_{T,K,p}, 1) & B(x_{T,K,p}, 2) & \dots & B(x_{T,K,p}, R) \end{bmatrix}$$

where $B(x, r)$ for $r = 1, \dots, R$ are basis functions; for our project, the B-spline basis was used. A recursive definition for the B-spline basis can also be found in Wood's book [Wood \(2017\)](#). Once the matrix $\Phi_{(2)}$ is defined, we QR decompose the vector $\Phi_{(2)}^T \mathbf{1}$ so that:

$$\Phi_{(2)}^T \mathbf{1} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}$$

Then, we select:

$$X_{(2)} = \Phi_{(2)} Q_2$$

Selecting the model matrix this way imposes an identifiability constraint, which states that the smooth function defined by the basis functions must satisfy:

$$\sum_{k=1}^K \sum_{t=1}^T f(x_{t,k,p}) = \mathbf{1}^T X_{(2)} \beta_{(2)} = 0$$

Notice that:

$$\mathbf{1}^T X_{(2)} \beta_{(2)} = \mathbf{1}^T \Phi_{(2)} Q_2 \beta_{(2)} = \begin{bmatrix} R & 0 \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} Q_2 \beta_{(2)} = R Q_1 Q_2 \beta_{(2)} = 0$$

since Q_1 and Q_2 are orthogonal. Simply stated, if $f(x)$ is a spline, the coefficients for the spline are now given by $Q_2 \beta_{(2)}$ instead of $\beta_{(2)}$ after the transformation. One more detail is that with the Poisson model, the likelihood:

$$\ln L = \sum_{k=1}^K \sum_{t=1}^T \left\{ \frac{y_{tk} \theta_{tk} - b(\theta_{tk})}{a(\phi)} + c(y_{tk}, \phi) \right\}$$

with $\theta_{tk} = \ln\{E[N_{tk}]\}$, is modified to:

$$\ln L^{\text{new}} = \sum_{k=1}^K \sum_{t=1}^T \left\{ \frac{y_{tk} \theta_{tk} - b(\theta_{tk})}{a(\phi)} + c(y_{tk}, \phi) \right\} + \lambda \beta_{(2)}^T Q_2^T S Q_2 \beta_{(2)}$$

where S is a matrix that penalizes the wiggleness of the function $f(x)$. In case the B-spline basis is used, difference penalties are imposed, as specified on page 206 of Wood's book [Wood \(2017\)](#). This procedure is implemented in the R library `mgcv`.

7. Results

In this section, we show the result of a simulation study using synthetically-generated data, similar to the real data that were analyzed during the project. Section [7.1](#) provides the details of the simulation study. In Section [7.2](#), we provide qualitative descriptions of the results obtained from the real data study.

7.1. Simulation Study

In order to perform a simulation study, synthetic data have been generated. The data have been generated according to the following rules to mimic the structure of the medical data, which we were unable to directly use for publication purposes due to a non-disclosure agreement.

- For each t , randomly generate the total number of patients from a Poisson distribution, so that $N_t \sim \text{Poisson} \left[5 \left\{ \sin \left(\frac{t}{365} \pi \right) + 2 \right\} \left\{ \sin \left(\frac{t}{365} 2\pi \right) + 2 \right\} \right]$, where t is the number of days elapsed since 1 January 2010, with the maximum t corresponding to 1 January 2019.
- Each patient i receives one treatment, whose ICD-10 code chapter is randomly generated from a multinomial distribution with the probability of each category following p_k , for $k = 1, \dots, 22$, sampled from a Dirichlet distribution.

- Department 1 opens at time 1 January 2017.
- Each treatment is assigned to either Department 1, 2, or 3 using a multinomial distribution with probability $(0, q_2, q_3)$, sampled from a Dirichlet distribution. We fixed $q_1 = 0$ before the opening of Department 1, because the probability that a treatment is assigned to Department 1 should be zero before its opening. We used a different set of probabilities (q'_1, q'_2, q'_3) , also sampled from a Dirichlet distribution, after Department 1 opens.
- Each treatment results in a positive charge with probability 0.95.
- Given that a charge in ICD-10 code chapter k is positive, it results in a charge amount sampled from a gamma distribution with its scale parameter sampled from an exponential distribution with rate 0.001 and shape parameters fixed to one.
- The total number of days in the synthetic data is 3287, with a total of 67,983 patients.

The modeling framework described in Section 5.1 has been used for the analysis of the synthetically-generated data. The regression coefficients are shown in Tables 3 and 4. From the regression coefficients for the number of patients, we see that the number of patients varied by month and chapter. From the charge severity models for Departments 1, 2, and 3, we see that the ICD-10 code chapters were a significant predictor of the charge amount. In Table 4, notice that the `ClinicOpen` variable was negative and significant for Department 2, indicating that the introduction of Department 1 has reduced the number of treatments given in Department 2. The time since 1 January 2010 has been used in conjunction with a smooth function constructed using the B-spline basis with order 10. Figure 1 shows a plot of the smooth function, which has been estimated using the GAM modeling approach described in Section 6. The plot illustrates that there is a nonlinear effect of the time variable on the number of patients.

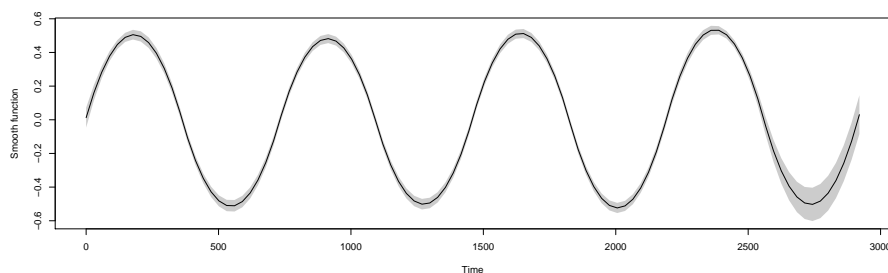


Figure 1. Fitted smooth function with respect to time.

Figure 2 shows the out-of-sample comparisons of the predicted expenditures and the synthetically-generated actual expenditures. Here, the out-of-sample data consisted of expenditures falling in the period between 1 January 2018 and 1 January 2019. Each panel shows the predictions when the result was aggregated at a daily, weekly, or monthly level. The Spearman correlation with the out-of-sample claims was 54.40%, 84.46%, and 95.78%, respectively. The Gini indices (explained more in Section 7.3) were 59.57%, 60.08%, and 62.37%, respectively.

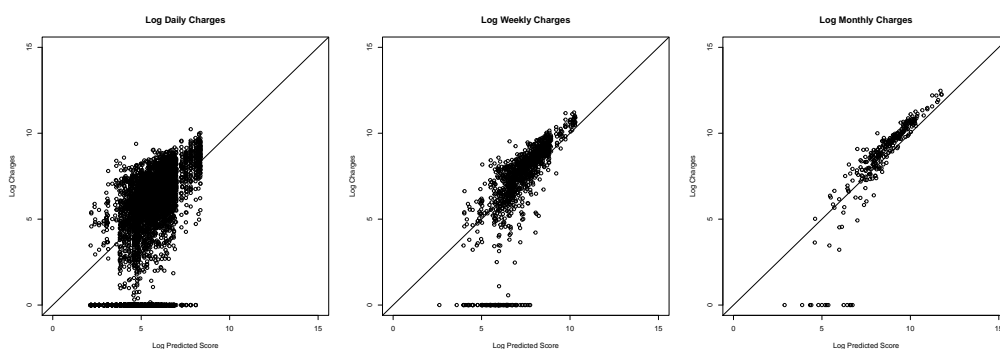


Figure 2. Log daily, weekly, and monthly charges (predicted versus actual out-of-sample charges).

Table 3. Model for number of patients and the treatments in Department 1.

Model for Number of Patients			
	<i>Estimate</i>	<i>Std. Err.</i>	
(Intercept)	−0.034	0.026	
ClinicOpen	0.017	0.050	
Month:2	0.131	0.019	***
Month:3	0.196	0.018	***
Month:4	0.180	0.018	***
Month:5	0.079	0.019	***
Month:6	−0.136	0.020	***
Month:7	−0.405	0.021	***
Month:8	−0.666	0.023	***
Month:9	−0.895	0.025	***
Month:10	−0.821	0.025	***
Month:11	−0.550	0.024	***
Month:12	−0.294	0.022	***
Chapter:2	0.824	0.024	***
Chapter:3	−1.869	0.055	***
Chapter:4	−0.278	0.031	***
Chapter:5	0.234	0.027	***
Chapter:6	−0.176	0.030	***
Chapter:7	0.531	0.025	***
Chapter:8	1.501	0.022	***
Chapter:9	0.357	0.026	***
Chapter:10	−1.801	0.053	***
Chapter:11	−1.217	0.042	***
Chapter:12	−1.288	0.043	***
Chapter:13	−0.800	0.036	***
Chapter:14	0.191	0.027	***
Chapter:15	−0.493	0.033	***
Chapter:16	0.269	0.027	***
Chapter:17	−0.494	0.033	***
Chapter:18	−1.584	0.049	***
Chapter:19	−2.909	0.088	***
Chapter:20	1.174	0.023	***
Chapter:21	−2.052	0.060	***
Chapter:22	0.078	0.028	**
Number of Treatments (Department 1)			
	<i>Estimate</i>	<i>Std. Err.</i>	
(Intercept)	−1.550	0.030	**
Probability of Positive Charge (Department 1)			
	<i>Estimate</i>	<i>Std. Err.</i>	
(Intercept)	0.952	0.004	***
Charge Severity Model (Department 1)			
	<i>Estimate</i>	<i>Std. Err.</i>	
(Intercept)	5.902	0.115	***
Chapter:2	−0.223	0.153	
Chapter:3	1.168	0.491	*
Chapter:4	1.309	0.216	***
Chapter:5	1.033	0.175	***
Chapter:6	0.936	0.223	***
Chapter:7	0.648	0.166	***
Chapter:8	−1.584	0.140	***
Chapter:9	−0.133	0.179	
Chapter:10	−1.271	0.394	**

Table 3. Cont.

Charge Severity Model (Department 1)			
	Estimate	Std. Err.	
Chapter:11	0.712	0.298	*
Chapter:12	0.282	0.419	
Chapter:13	0.917	0.232	***
Chapter:14	0.963	0.193	***
Chapter:15	−0.246	0.229	
Chapter:17	0.741	0.223	***
Chapter:18	1.197	0.419	**
Chapter:19	1.710	1.073	
Chapter:20	1.399	0.145	***
Chapter:21	1.097	0.491	*
Chapter:22	−0.266	0.201	
Significance: '***': 0.001, '**': 0.01, '*': 0.05, '.': 0.1			

7.2. Real Data Analysis

In this section, we present a qualitative analysis of the results from the treatment-level medical data modeling using real data. Because of the non-disclosure agreement, we provide qualitative results only. The model for the total number of patients N_{tk} showed that `WDay`, `Month`, `Chapter`, and `ClinicOpen` (all explanatory variables) were statistically significant. There was also a nonlinear relationship between the number of patients arriving and time, which could be inferred from the GAM approach to using the `Time` variable as the independent variable for a smooth function. The coefficient for the variable `ClinicOpen` could be used to infer the relationship between the total number of patients and the opening of Department 1. The number of treatments model for $M_{2,tki}$ and $M_{3,tki}$ provided interesting results. The `ClinicOpen` variable indicated that the number of treatments in the other two departments changed following the opening of Department 1.

7.3. Model Validation

After estimating all the parameters for the hierarchical model, the model can be validated using an external test set—commonly called an “out-of-sample” comparison—where a subset of the initial dataset observations was set aside. In our analysis, the subset of observations from a particular year was the out-of-sample set (test set), and the remaining samples (training set) were used to estimate the parameters (train the model). The predicted values of the out-of-sample observations determined using the model were compared with their observed values. A graphical approach plots the observed out-of-sample expenditures with their predicted expenditures to visualize how close the points fell near the 45-degree identity line. The Spearman correlation between the predicted expenditures and the out-of-sample expenditures is a good initial method to measure whether the ranks of the observations were predicted well. A new and more informative measure was developed by [Frees et al. \(2011\)](#) employing the Gini index. The Gini index is defined as twice the area underneath the ordered Lorenz curve. In order to obtain the ordered Lorenz curve, let:

$$\hat{F}_{\Pi}(s) = \frac{\sum_{t=1}^T \sum_{k=1}^K \Pi(\mathbf{x}_{tk}) I(R(\mathbf{x}_{tk}) \leq s)}{\sum_{t=1}^T \sum_{k=1}^K \Pi(\mathbf{x}_{tk})} \quad \text{and} \quad \hat{F}_L(s) = \frac{\sum_{t=1}^T \sum_{k=1}^K y_{tk} I(R(\mathbf{x}_{tk}) \leq s)}{\sum_{t=1}^T \sum_{k=1}^K y_{tk}},$$

where we define the relativity $R(\mathbf{x}_{tk})$ as:

$$R(\mathbf{x}_{tk}) = \frac{S(\mathbf{x}_{tk})}{\Pi(\mathbf{x}_{tk})}, \quad S(\mathbf{x}_{tk}) = \text{insurance score}, \quad \Pi(\mathbf{x}_{tk}) = \text{constant score}.$$

The ordered Lorenz curve can be obtained by plotting the points:

$$(\hat{F}_{\Pi}(s), \hat{F}_L(s)).$$

The constant charge model assumes that the charge for a day in every treatment category is the average cost over all treatment-days for the in-sample (training set) subset of observations. Our model validation approaches demonstrated that our model had a positive and significant Gini index.

7.4. Dependence Modeling

The Spearman correlation is a measure of dependency among the ranks of random variables. We tested the Spearman correlation among the number of treatments $M_{1,tki}$, $M_{2,tki}$, and $M_{3,tki}$ and discovered that there was evidence of negative dependence. Because there was evidence of dependence, a potentially interesting avenue of future work is the application of copula models to the discrete number of treatment variables. For an introduction to copula models, the reader is referred to Nelsen's introduction to copulas [Nelsen \(1999\)](#), while [Joe \(2014\)](#) presented an in-depth monograph for the use of copulas to model complex dependence structures. One final consideration is the validity of the independence assumption between the number of patients N_{tk} , the number of treatments $M_{1,tki}$, $M_{2,tki}$, $M_{3,tki}$, the positivity of the treatment charges $P_{1,tkij}$, $P_{2,tkij}$, $P_{3,tkij}$, and the charge amounts $Y_{1,tkij}$, $Y_{2,tkij}$, $Y_{3,tkij}$. We emphasize that it is an important assumption to treat these variables to be independent in order for our modeling approach to be valid.

Table 4. Model for the treatments at Departments 2 and 3.

Number of treatments (Department 2)			
	Estimate	Std. Err.	
(Intercept)	−0.370	0.005	***
ClinicOpen	−0.451	0.022	**
Probability of Positive Charge (Department 2)			
	Estimate	Std. Err.	
(Intercept)	0.949	0.001	**
Charge Severity Model (Department 2)			
	Estimate	Std. Err.	
(Intercept)	6.826	0.020	***
Chapter:2	−1.121	0.026	***
Chapter:3	0.199	0.067	**
Chapter:4	0.436	0.035	***
Chapter:5	0.049	0.030	
Chapter:6	0.233	0.034	***
Chapter:7	−0.392	0.028	***
Chapter:8	−2.479	0.023	***
Chapter:9	−1.212	0.029	***
Chapter:10	−1.759	0.067	***
Chapter:11	−0.236	0.050	***
Chapter:12	−0.835	0.052	***
Chapter:13	−0.433	0.043	***
Chapter:14	0.004	0.031	
Chapter:15	−1.411	0.038	***
Chapter:16	−2.991	0.030	***
Chapter:18	0.436	0.059	***
Chapter:19	0.184	0.110	.
Chapter:20	0.491	0.024	***
Chapter:21	−0.014	0.071	
Chapter:22	−1.393	0.031	***

Table 4. Cont.

Number of Treatments (Department 3)			
	Estimate	Std. Err.	
(Intercept)	−1.094	0.007	***
ClinicOpen	0.079	0.024	**
Probability of Positive Charge (Department 3)			
	Estimate	Std. Err.	
(Intercept)	0.948	0.002	***
Charge Severity Model (Department 3)			
	Estimate	Std. Err.	
(Intercept)	6.625	0.036	***
Chapter:2	−0.928	0.043	***
Chapter:3	0.456	0.101	***
Chapter:4	0.636	0.054	***
Chapter:5	0.244	0.048	***
Chapter:6	0.374	0.053	***
Chapter:7	−0.186	0.045	***
Chapter:8	−2.267	0.039	***
Chapter:9	−0.978	0.047	***
Chapter:10	−1.469	0.092	***
Chapter:11	−0.019	0.075	
Chapter:12	−0.532	0.077	***
Chapter:13	−0.069	0.063	
Chapter:14	0.206	0.048	***
Chapter:15	−1.152	0.058	***
Chapter:16	−2.827	0.048	***
Chapter:17	0.404	0.058	***
Chapter:18	0.671	0.087	***
Chapter:19	0.373	0.157	*
Chapter:20	0.682	0.041	***
Chapter:21	−0.124	0.113	
Chapter:22	−1.150	0.050	***

Significance: '***': 0.001, '**': 0.01, '*': 0.05, '.': 0.1

8. Conclusions

In this paper, we have presented two unique analysis and modeling endeavors, (1) store-level sales modeling project and (2) a medical treatment-healthcare expenditure modeling project. The store sales project demonstrated the need to combine nearby business with residential demographics information. While the store sales project is not a traditional actuarial science project, it demonstrated to the students the wide applicability of the tools and methods they have learned during their undergraduate career. The store sales project also demonstrated the power (and need) to identify and harvest data outside of the provided dataset to produce informative predictive models.

The medical treatment-healthcare expenditure project was a more traditional actuarial science project and introduced the undergraduate students to real-world and messy medical data. The treatment-level healthcare data resulted in expenditure models that were motivated by the traditional hierarchical insurance claims models typically found in actuarial science. The healthcare cost models provided insights into the effect of the explanatory variables on the component response variables. In a way, the store-level sales modeling was a general introduction to generalized linear models that lead into frequency-severity and hierarchical models.

The presented research provides an overview of our generalized philosophy on modeling and analyzing healthcare expenditures that includes the creation of informative and predictive models to help understand the provided data, answer questions about the system of interest, and provide

insights, all while ensuring the models are robust and predictive. The store sales models indicated that the store's total size and the population count within certain miles of the stores are key features driving store sales for the year of interest. For the medical treatment analysis, our research indicated that the overall number of patients have a nonlinear relationship with the time variable, and this changed after the introduction of Department 1 into the healthcare network. The number of treatments for a given patient also changed in Departments 2 and 3 after the opening of Department 1 to the public. These analyses give insights into the impact of the explanatory variables on the response variables of interest. In addition, we demonstrated a treatment-level approach to estimating the total expenditures incurred to a healthcare system by patients.

We believe that our approach can be useful for predicting healthcare expenditures for the healthcare services sector. The approach can be applied to any healthcare services network with multiple departments and treatment-level records of patients. More information at the patient level (for example, demographic information) may improve the model. Because our model has the ability to predict the number of patients arriving at each department under each category of treatment, we believe that the modeling approach can help efficiently allocate resources (including physician times) at each healthcare department. We believe that this can contribute to improving the quality of healthcare services provided to the patients.

9. Disclaimer

This paper is the result of a project that took place in the form of an undergraduate Teamwork Experience course (MTH491B) at Michigan State University (MSU) during the fall semester of 2017 and continued on to the 2018 spring semester. A non-profit healthcare provider approached the Actuarial Science department of MSU with a treatment-level dataset. Their request was for us to analyze and provide insights to aid their operation. The MSU team consisted of three undergraduate students, a graduate student in the Department of Statistics and Probability Ph.D. program, an Assistant Professor with a joint appointment in the Department of Statistics and Probability and the Department of Mathematics, and a professional predictive analytics modeling consultant. The project initially focused on discovering treatment-and patient-level characteristics of the data and creating summary statistics of the data, along with predicting the total amount of cost for the operation of the healthcare provider. Through the application of statistical analysis and models, the goal was to uncover new aspects of the data for the non-profit's management team. In addition to the medical data, the non-profit healthcare provider supplied an additional and unrelated dataset of stores containing the sales volume and store features maintained within the non-profit's network. As there were two datasets to analyze, the project consisted of two components, one being the analysis of the store data, and the other being the treatment-level analysis. The store sales analysis portion of the project turned into a University Undergraduate Research and Arts Forum (UURAF) presentation at MSU by the undergraduate students involved in the project at the end of the spring 2018 semester. The title of the poster presentation was "Combining Population and Nearby Store Information to Aid in Selection of Future Store Locations." In this paper, we have discussed both components of the project with the goal of reviewing the general approaches that could be used in order to analyze healthcare data at the store level and the treatment level.

Author Contributions: All authors contributed substantially to this work.

Funding: This research received no external funding.

Acknowledgments: This work was made possible by the generous support and resources provided by an anonymous non-profit healthcare provider in Michigan. The authors are grateful for their support throughout the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Boucher, Jean-Philippe. 2014. Regression with count dependent variables. In *Predictive Modeling Applications in Actuarial Science*. Cambridge: Cambridge University Press.
- De Jong, Piet, and Gillian Z. Heller. 2008. *Generalized Linear Models for Insurance Data*. Cambridge: Cambridge University Press.
- Dobson, Annette J., and Barnett Adrian G. 2008. *An Introduction to Generalized Linear Models*. Boca Raton: Chapman and Hall/CRC, Taylor & Francis Group.
- Frees, Edward W. 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.
- Frees, Edward W. 2009. *Regression Modeling with Actuarial and Financial Applications*. Cambridge: Cambridge University Press.
- Frees, Edward W. 2014. Frequency and severity models. In *Predictive Modeling Applications in Actuarial Science*. Cambridge: Cambridge University Press.
- Frees, Edward W. 2015. Analytics of insurance markets. *Annual Review of Financial Economics* 7: 253–77. [[CrossRef](#)]
- Frees, Edward W., Jie Gao, and Marjorie A. Rosenberg. 2011. Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal* 15: 377–92. [[CrossRef](#)]
- Frees, Edward W., and Gee Lee. 2016. Rating endorsements using generalized linear models. *Variance* 10: 51–74.
- Frees, Edward W., Gee Y. Lee, and Lu Yang. 2016. Multivariate frequency-severity regression models in insurance. *Risks* 4: 4. [[CrossRef](#)]
- Frees, Edward W., Glenn Meyers, and A. David Cummings. 2011. Summarizing insurance scores using a gini index. *Journal of the American Statistical Association* 106: 495. [[CrossRef](#)]
- Frees, Edward W., and Emiliano A. Valdez. 2008. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103: 1457–69. [[CrossRef](#)]
- Guillén, Montserrat. 2014. Regression with categorical dependent variables. In *Predictive Modeling Applications in Actuarial Science*. Cambridge: Cambridge University Press.
- Joe, Harry. 2014. *Dependence Modeling with Copulas*. Boca Raton: CRC Press.
- Keeler, Emmett B., and John E. Rolph. 1988. The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics* 7: 337–67. [[CrossRef](#)]
- Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. 2012. *Loss Models: From Data to Decisions*. Hoboken: John Wiley & Sons, Inc.
- Mildenhall, Stephen J. 1999. A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society* 86: 393–487.
- Myers, Raymond, Douglas C. Montgomery, G. Geoffrey Vining, and Timothy J. Robinson. 2002. *Generalized Linear Models with Applications in Engineering and the Sciences*. New York: John Wiley & Sons, Inc.
- Nelder, John, and Robert Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135: 370–84. [[CrossRef](#)]
- Nelsen, Roger B. 1999. *An Introduction to Copulas*. New York: Springer Science & Business Media, Inc.
- Ohlsson, Esbjörn, and Björn Johansson. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin Heidelberg: Springer Verlag.
- Rosenberg, Marjorie A., and Phillip M. Farrell. 2008. Predictive modeling of costs for a chronic disease with acute high-cost episodes. *North American Actuarial Journal* 12: 1–19. [[CrossRef](#)]
- Ruscone, Marta Nai, and Silvia Angela Osmetti. 2016. *Modelling the Dependence in Multivariate Longitudinal Data by Pair Copula Decomposition*. Basel: Springer International Publishing Switzerland.
- Shi, Peng. 2012. Multivariate longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics* 51: 204–15. [[CrossRef](#)]
- Shi, Peng. 2014. Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science*. Cambridge: Cambridge University Press.
- Shi, Peng, and Emiliano Valdez. 2014. Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal* 2014: 159–79. [[CrossRef](#)]
- Smith, Michael, Aleksey Min, Carlos Almeida, and Claudia Czado. 2010. Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association* 105: 1467–79. [[CrossRef](#)]

- Sun, Jiafeng, Edward W. Frees, and Marjorie A. Rosenberg. 2008. Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* 42: 817–30. [[CrossRef](#)]
- Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R, Second Edition*. Boca Raon: CRC Press.
- Yang, Xipei. 2011. Multivariate Long-Tailed Regression With New Copulas. Ph.D. thesis, University of Wisconsin-Madison, Madison, WI, USA.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).