August 2023

Excess zeros under GAM: Tweedie or two-part?

Xianming Zeng
University of Windsor

Excess zeros under GAM: Tweedie or two-part?

By

Xianming Zeng

A Major Research Paper
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2023

Excess zeros under GAM: Tweedie or two-part?

by

Xianming Zeng

APPROVED BY:

K. Granville
Department of Mathematics and Statistics

A. Hussein, Co-Advisor
Department of Mathematics and Statistics

M.Belalia, Co-Advisor
Department of Mathematics and Statistics

June 27, 2023

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this major paper and that no part of this major paper has been published or submitted for publication.

I certify that, to the best of my knowledge, my major paper does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my major paper, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my major paper and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my major paper, including any final revisions, as approved by my major paper committee and the Graduate Studies office, and that this major paper has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Positive, right-skewed data with excess zeros are encountered in many real-life situations. Two possible techniques to analyze this type of data are: Two-part models and Tweedie models. The two-part models assume existence of a separate zero-generating process, while the Tweedie models are based on distributions that allow mass at zero. The paper aims to present a simulation study to investigate the performance of Generalized Additive Models (GAM) under the distribution of Tweedie and two-part models for such data with excess zero by using MSE (Mean Square Error) and relative bias to compare the performance of both methods. We found that under different practical scenarios, the two-part model has a better performance than the Tweedie.

DEDICATION

This dedication is to all those who have supported and encouraged me throughout my journey. To my family, friends, mentors, and loved ones - thank you for believing in me and for always being there to offer guidance and encouragement.

To my kind mentor who inspired me - your creativity, passion, and resilience have shown me what is possible and have motivated me to keep pushing forward.

To my classmates who challenged me - thank you for pushing me out of my comfort zone and helping me grow.

To my dad who have taught me a lot in both my life and study- your knowledge, wisdom, and experience have been invaluable in shaping who I am today.

And to all those who have touched my life in some way - thank you for being a part of my journey and for helping me become the person I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

The analysis of zero-inflated data as illustrated by Tu (2006) is a challenging problem in statistical modeling, and has become increasingly important in many fields, including agricultural (Hall, 2000), medical (Böhning et al., 1999), manufacturing (Lambert, 1992) and economics (Freund et al., 1999). Zero-inflated continuous data is a type of data where there is an excess number of zero values in a continuous variable, beyond what would be expected from a reference continuous distribution, see Liu et al. (2019) as an example. This kind of data are difficult to model using standard regression models that assume a continuous distribution, such as the Normal distribution or Gamma distribution. This is because the excessive number of zeros can lead to skewness and non-normality, which can violate the assumptions of the model. The objective of this study is to explore the performance of the GAM (Generalized Additive Models) by using two different methods when there are excess zeros in the data.

The GAM, which were proposed by Hastie (2017) are flexible and powerful class of models that can be used to analyze complex relationships between response and predictor variables in a wide range of fields. The two popular approaches for modeling zero-inflated data are Tweedie distribution, see Shono (2008) as an example and Two-part model, see Duan et al. (1983) as an example. The Tweedie distribution is a flexible family of distributions that can accommodate both overdispersion and underdispersion, while the two-part model assumes the existence of a zero-generating process (a distribution with mass at zero) and continuous distribution that generates the non-zero part of the data. In this paper, we present a simulation study

to evaluate the performance of these two approaches under the GAM model. We set the proportion of zero data into four different levels, roughly 5%-15%, 15%-25%, 25%-35% and 35%-45%. In addition we generate the independent variable from a uniform distribution and the continuous part of the dependent variable is set to follow a Gamma distribution with 2 parameters that depend on a set of covariates through some known functions. We use Mean Squared Error, MSE, and integrated percent bias to evaluate the performance of the two approached of accommodating the zeros (Tweedie and two-part). According to the result we will provide recommendations for selecting appropriate models for analyzing zero-inflated data using GAM. Our results show that GAM with the two-part model has better performance and can be a useful tool for analyzing zero-inflated data, especially when the true distribution of the data is unknown or cannot be modeled with traditional regression models.

The rest of the paper is structured as follows: chapter 2 will go through some background of GAM, Tweedie distribution, and two-part model. Chapter 3 outlines the simulation process and how the generated data and functions look like, and compare the final result to give a conclusion and discuss some further works that can be done. Chapter 4 gives the conclusion and some future work that can be done.

CHAPTER 2

Background of modelling methodologies

In this chapter, we will go through some concepts of GAM, Tweedie and two-part model before we model the zero-inflated data. Section 2.1 will be talking about the definition and properties of Tweedie distributions, Section 2.2 will include the GAM properties and the cubic spline regression smoothing method. Section 2.3 demonstrates the working mechanism of two-part model using Gamma distribution.

2.1 Tweedie distributions

In probability and statistics, the Tweedie distributions are defined as a family of power variance functions which include the purely discrete case like Poisson distribution and the Poisson-gamma distribution with a positive mass at zero and otherwise positively continuous. The class was first introduced by Tweedie et al. (1984) and then named and classified by Jorgensen (1997). Now Tweedie distributions are a special case of exponential dispersion models and widely applied in many areas like ecology (Foster and Bravington, 2013), insurance (Shi, 2016) and fisheries research (Candy, 2004). Basically when the data is a mixture of zeros and non-negative points as shown in Figure 2.1.1, it is possible to fit the data using Tweedie distribution.

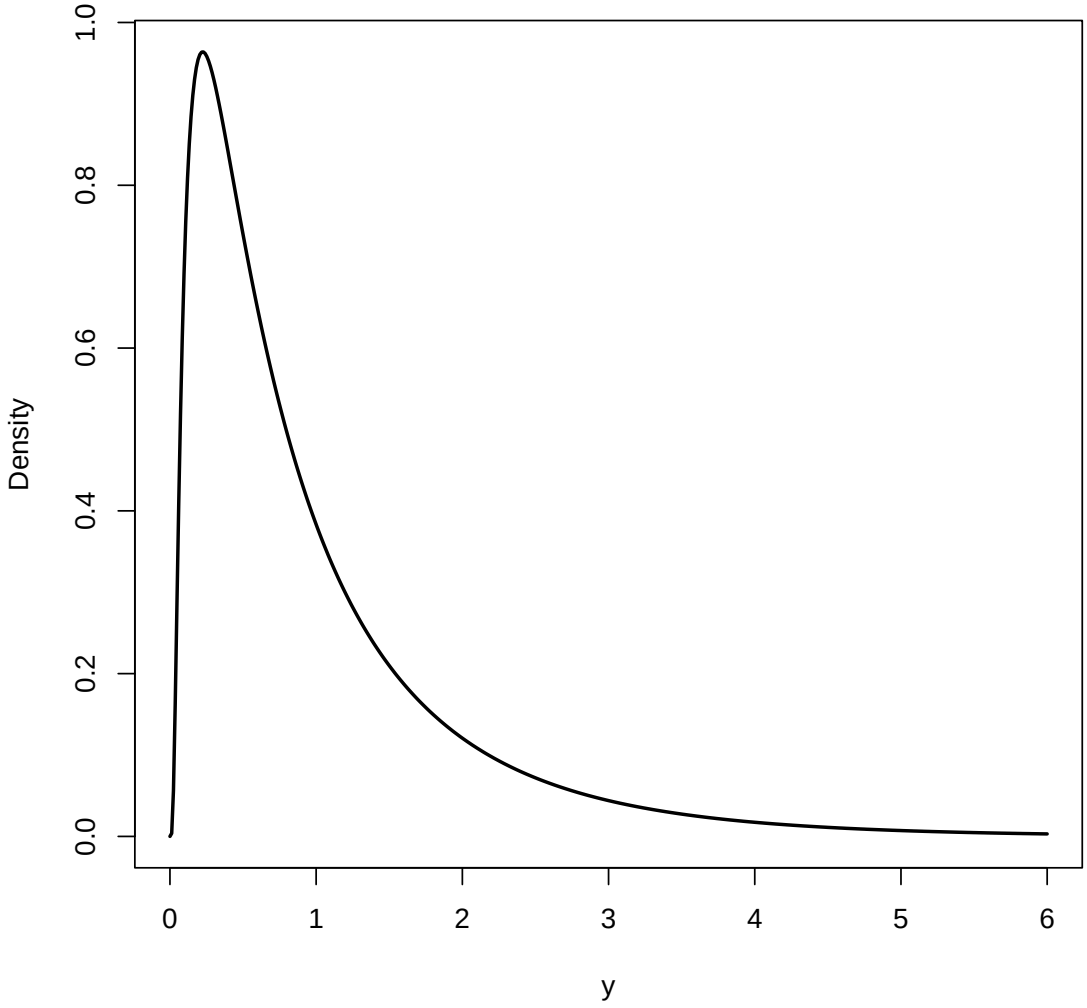

Fig. 2.1.1: An example of Tweedie distribution showing the point mass at zero

The Tweedie distributions, which are used to model reproductive data, are a subset of exponential dispersion models (ED) defined by Jorgensen (1987) which exhibit a unique mean-variance relationship.

Definition 1. *Following Jorgensen (1997), Let C be a convex support and let Ω be the interior of C. They are intervals satisfying: $\Omega \subseteq C \subseteq R$. A unit deviance is a function $d : C \times \Omega \to R$ which satisfies the following properties:*

(i). $d(y, y) = 0$, for $\forall y \in \Omega$;

(ii). $d(y, x) > 0$, for $\forall y \neq x$.

A random variable is said to follow the Tweedie distribution $T_{W_p}(\mu, \sigma^2)$ if $Y \sim ED(\mu, \sigma^2)$ with expectation $E(Y) = \mu$. The class of EDMs (Exponential dispersion models) is a family which has the axiomatic definition given by:

Definition 2. *An Exponential Dispersion Model (EDM) is a probability distribution whose pdf is defined as:*

$$f(y; \mu, \sigma^2) = a(y, \sigma^2) \exp\left(-\frac{1}{2\sigma^2} d(y, \mu)\right), \quad y \in C,$$

where $a \geq 0$ is a suitable function, d is a unit deviance of the form $d(y, \mu) = yg(\mu) + h(\mu) + k(y)$, C is the convex support, $\mu \in \Omega$, and Ω is an interval.

Here we consider the function $a(\cdot)$ is suitable if the condition of probability that the integration of pdf equals to 1 is satisfies. In other words, it is sufficient to check:

$$\int f(y; \mu, \sigma^2) dy = 1. \tag{2.1.1}$$

The terminology "dispersion model" is named for interpreting the $\sigma^2 > 0$ as the dispersion parameter. For fixed parameter $\sigma^2 > 0$, the $ED(\mu, \sigma^2)$ is just the natural exponential family. μ is the position parameter and the expectation of Y is given by $E(Y) = \mu$, the variance is given by $Var(Y) = \sigma^2 V(\mu)$, where $V(\cdot)$ is called the unit variance function.

As was illustrated in Jorgensen (1997), we would elaborate the constructive definition form for EDMs.

Definition 3. *Let v be a σ-finite measure on $\mathbb{R}$. The cumulant function $\kappa(\theta)$ is defined as:*

$$\kappa(\theta) = \log \int e^{\theta y} v(dy).$$

The domain of $\kappa(\theta)$ is $\Theta = \left\{\theta \in \mathbb{R} : \left(\int e^{\theta y} v(dy) < \infty\right)\right\}$.

In practice, we rarely compute the cumulant function. There are other convenient ways of defining the cumulant functions. We will discuss this later, but right now we will define the measure v. Note that Both constructive definition and axiomatic

definition give the same idea about the EDMs. Before we show the proof of this statement, we need some preparations. Recall that we mentioned $a(y, \sigma^2)$ is any suitable function satisfying the Equation (2.1.1), in general, $a(y, \sigma^2)$ does not have closed form. Let $b(y, \sigma^2)$ be a function like $a(y, \sigma^2)$, now we take $\sigma^2 = 1$, we can get that:

$$v(dy) = b(y, 1)dy. \tag{2.1.2}$$

Here dy is the Lebesgue measure. For any random variable Y which has one parameter θ and defined on a measurable set A, the cdf is:

$$P_\theta(Y \in A) = \int_A \exp\{y\theta - \kappa(\theta)\}\, v(dy). \tag{2.1.3}$$

By substitute Expression (2.1.2) in Expression (2.1.3), we can have that it is equivalent to the following:

$$\int_A \exp\{y\theta - \kappa(\theta)\}\, b(y, 1)dy. \tag{2.1.4}$$

Now we can determine the MGF (Moment Generating Function) and the CGF (Cumulant Generating Function) of the random variable Y. The MGF of random variable Y is defined with Expression (2.1.3) as:

$$\begin{aligned} M_Y(t; \theta) &= \int \exp\{yt\} \exp\{y\theta - \kappa(\theta)\} v(dy) \\ &= \exp\{-\kappa(\theta)\} \int \exp\{yt + y\theta\} v(dy) \\ &= \exp\{-\kappa(\theta)\} \exp\{\kappa(\theta + t)\} \\ &= \exp\{\kappa(\theta + t) - \kappa(\theta)\}. \end{aligned}$$

Now we get the MGF of Y, the CGF is just the log of the MGF, thus, we can easily obtain that the CGF is:

$$K_Y(t; \theta) = \kappa(\theta + t) - \kappa(\theta).$$

By the property of CGF, we know that the first derivative of CGF computed at $t = 0$, is the mean μ, and the second derivative computed at $t = 0$ is the variance of the random variable, then we can observe that:

$$K^{(i)}(t;\theta) = \frac{\partial^{(i)} K(t;\theta)}{\partial t^i} = \kappa^{(i)}(\theta + t)$$
$$K^{(i)}(0;\theta) = \kappa^{(i)}(\theta).$$

Now it is trivial to see that:

$$\kappa'(\theta) = \mu.$$

As in the EDMs, we see that there are 2 parameters of μ and σ^2, we can induce from the definition for one-parameter case and give the definition for random variable Y in exponential family with two parameters as:

$$P_{\theta,\sigma^2}(Y \in A) = \int_A \exp\left(\frac{y\theta - \kappa(\theta)}{\sigma^2}\right) v(dy). \tag{2.1.5}$$

Thus, we can give the constructive definition of EDMs as:

Definition 4. *An Exponential Dispersion Model (EDM) is a probability distribution whose pdf is defined as:*

$$f(y;\mu,\sigma^2) = b(y,\sigma^2)\exp\left(\frac{y\theta - \kappa(\theta)}{\sigma^2}\right), \quad y \in C,$$

where $b \geq 0$ is some suitable function and C is a convex support, $\mu \in \Omega$ as well.

By applying a similar calculation process as shown above, we can compute the

MGF and CGF for the two-parameter case as:

$$\begin{aligned} M_Y(t;\theta,\sigma^2) &= \int \exp\{yt\} \exp\left\{\frac{y\theta-\kappa(\theta)}{\sigma^2}\right\} v(dy) \\ &= \exp\left\{-\frac{\kappa(\theta)}{\sigma^2}\right\} \int \exp\left\{yt+\frac{y\theta}{\sigma^2}\right\} v(dy) \\ &= \exp\left\{-\frac{\kappa(\theta)}{\sigma^2}\right\} \exp\left\{\frac{\kappa(\theta+t\sigma^2)}{\sigma^2}\right\} \\ &= \exp\left\{\frac{\kappa(\theta+t\sigma^2)-\kappa(\theta)}{\sigma^2}\right\}, \end{aligned}$$

and

$$K_Y(t;\theta,\sigma^2) = \frac{\kappa(\theta+t\sigma^2)-\kappa(\theta)}{\sigma^2}. \tag{2.1.6}$$

Before we give the proof that both constructive definition and axiomatic definition are equivalent, we need a function that maps from parameter space Θ to position parameter space Ω in EDMs. Define a function $\tau : \tau(\theta) = \kappa'(\theta) = \mu$, $\Theta \to \Omega$ where Θ is the parameter space, and Ω is the mean parameter space. Let τ^{-1}:$\tau^{-1}(\mu) = \theta$, $\Omega \to \Theta$ be the inverse function of τ. Taking the first and second derivative of Equation (2.1.6) with respect to t at $t = 0$, we can find that:

$$K'_Y(0;\theta,\sigma^2) = \frac{\sigma^2\kappa'(\theta)}{\sigma^2} = \kappa'(\theta) = \tau(\theta) = \mu \tag{2.1.7}$$

$$K''_Y(0;\theta,\sigma^2) = \frac{(\sigma^2)^2\kappa''(\theta)}{\sigma^2} = \sigma^2\kappa''(\theta) = \sigma^2\tau'(\theta). \tag{2.1.8}$$

Recall that the variance of EDMs is $\sigma^2 V(\mu)$, therefore, we can know that $\tau'(\theta) = \kappa''(\theta) = V(\mu)$. Now from the cdf defined in Expression (2.1.5), we can rewrite the pdf, after setting the σ-finite measure v to be $v = b(y;\sigma^2)dy$,

$$f(y;\mu,\sigma^2) = b(y;\sigma^2)\exp\left\{\frac{y\theta-\kappa(\theta)}{\sigma^2}\right\} = b(y;\sigma^2)e^{\frac{y\tau^{-1}(\mu)-\kappa(\tau^{-1}(\mu))}{\sigma^2}}. \tag{2.1.9}$$

As was proposed in Jorgensen (1997), he gave the unit deviance $d(y;\mu)$ for the

exponential dispersion model as:

$$2\left[\sup_{\theta\in\Theta}(y\theta-\kappa(\theta)-y\tau^{-1}(\mu)+\kappa(\tau^{-1}(\mu))\right].$$

Taking partial derivative with respect to θ to find the maximum of the above form, we can get that:

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(y\theta-\kappa(\theta)\right)=y-\kappa'(\theta)=0\\ \Rightarrow y-\tau(\theta)=0\\ \Rightarrow\theta=\tau^{-1}(y).\end{aligned}$$

Thus, we can get the unit deviance $d(y;\mu)$ for $y\in\Omega$ as:

$$2\left[y\tau^{-1}(y)-\kappa\left(\tau^{-1}(y)\right)-y\tau^{-1}(\mu)+\kappa\left(\tau^{-1}(\mu)\right)\right]. \tag{2.1.10}$$

By using Expression (2.1.10), in the following proposition we show that both axiomatic definition and constructive definition for EDMs are the same.

Proposition 1. *Let f_a be the pdf of axiomatic definition and f_c is the pdf of constructive definition. Suppose that $a(y;\sigma^2)=f_c(y;y,\sigma^2)$, then we have that:*

$$f_a(y;\mu,\sigma^2)=f_c(y;\mu,\sigma^2).$$

Proof. By Expression (2.1.10), we can have that for $y\in\Omega$, the unit deviance is:

$$d(y,\mu)=2\left[y\tau^{-1}(y)-\kappa\left(\tau^{-1}(y)\right)-y\tau^{-1}(\mu)+\kappa\left(\tau^{-1}(\mu)\right)\right].$$

Substituting this into Definition 2, we have that:

$$\begin{aligned} f_a(y;\mu,\sigma^2) &= a(y,\sigma^2)\exp\left(-\frac{1}{2\sigma^2}d(y,\mu)\right) \\ &= f_c(y;y,\sigma^2)\exp\left(-\frac{1}{2\sigma^2}d(y,\mu)\right) \\ &= b(y;\sigma^2)\exp\left\{\frac{y\tau^{-1}(y)-\kappa(\tau^{-1}(y))}{\sigma^2}\right\}\exp\left(-\frac{1}{2\sigma^2}d(y,\mu)\right) \\ &= b(y;\sigma^2)\exp\left\{\frac{y\tau^{-1}(\mu)-\kappa(\tau^{-1}(\mu))}{\sigma^2}\right\} \\ &= b(y;\sigma^2)\exp\left\{\frac{y\theta-\kappa(\theta)}{\sigma^2}\right\} \\ &= f_c(y;\mu,\sigma^2). \end{aligned}$$

□

Here we complete the proof and say that both definitions are the same on convex supports, which are the most common cases. The proof for the general case of this idea can be found in the book of Jorgensen (1997).

Definition 5. *The Tweedie family of distributions $Tw_p(\mu,\sigma^2)$ is a special case of EDMs where the power mean-variance relationship is characterized by*

$$Var(Y) = \sigma^2\mu^p, \tag{2.1.11}$$

where $p \in (-\infty,0]\cup[1,\infty)$ and $\sigma^2 > 0$, p is called the Tweedie power parameter.

This expression indicates that the unit variance function is $V(\mu) = \mu^p$, thus, we can observe that:

$$\kappa''(\theta) = \tau'(\theta) = \frac{\partial\mu}{\partial\theta} = \mu^p.$$

From Equation (2.1.11), we can determine the expression for parameter θ in terms of μ and p. Ignoring the constant term, we can get that:

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1 \\ \log(\mu) & p = 1. \end{cases}$$

Next, we introduce a parameter α that is only related to p to compute μ in terms of θ and α, the relationship is defined as below:

$$\alpha = \frac{p-2}{p-1} \quad \text{or} \quad p = \frac{\alpha-2}{\alpha-1}.$$

The μ represented in terms of θ and α is:

$$\mu = \begin{cases} \left(\frac{\theta}{\alpha-1}\right)^{\alpha-1} & p \neq 1 \\ e^{\theta} & p = 1. \end{cases} \tag{2.1.12}$$

Using Equation (2.1.12), we can find the cumulant function $\kappa(\theta)$ by solving the differential equation $\kappa'(\theta) = \tau(\theta) = \mu$ for $p \neq 1, 2$, we can get that

$$\kappa(\theta) = \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha}.$$

For $p = 1$, we have $\kappa(\theta) = e^{\theta}$. When $p = 2$, we have that $\alpha = 0$, which implies that $\kappa'(\theta) = \frac{-1}{\theta}$. Taking the anti-derivative we get that $\kappa(\theta) = -\log(-\theta)$ when $p = 2$. In summary, we have that in Tweedie distribution, the cumulant function $\kappa(\theta)$ is defined as:

$$\kappa(\theta) = \begin{cases} \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha} & \text{for } p \neq 1, 2 \\ -\log\left(-\theta\right) & \text{for } p = 2 \\ e^{\theta} & \text{for } p = 1. \end{cases} \tag{2.1.13}$$

This is a simple analytic way of computing the cumulant generating function. To have an idea about the behaviour of the Tweedie distributions when we vary these parameters, Table 2.1.1 shows the Tweedie models based on indexing parameter p, and Figure 2.1.2 shows how the Tweedie distribution looks like when μ and σ^2 change.

Table 2.1.1: Tweedie models based on different indexing parameter p

Distribution	p	Domain	Mean domain
Stable	$p < 0$	$\mathbb{R}$	$(0, \infty)$
Normal	$p = 0$	$\mathbb{R}$	$\mathbb{R}$
Do not exist	$0 < p < 1$		
Poisson	$p = 1$	$\mathbb{N}$	$(0, \infty)$
Compound Poisson-gamma	$1 < p < 2$	$[0, \infty)$	$(0, \infty)$
Gamma	$p = 2$	$(0, \infty)$	$(0, \infty)$
Stable	$2 < p < 3$	$(0, \infty)$	$(0, \infty)$
Inverse Gaussian	$p = 3$	$(0, \infty)$	$(0, \infty)$

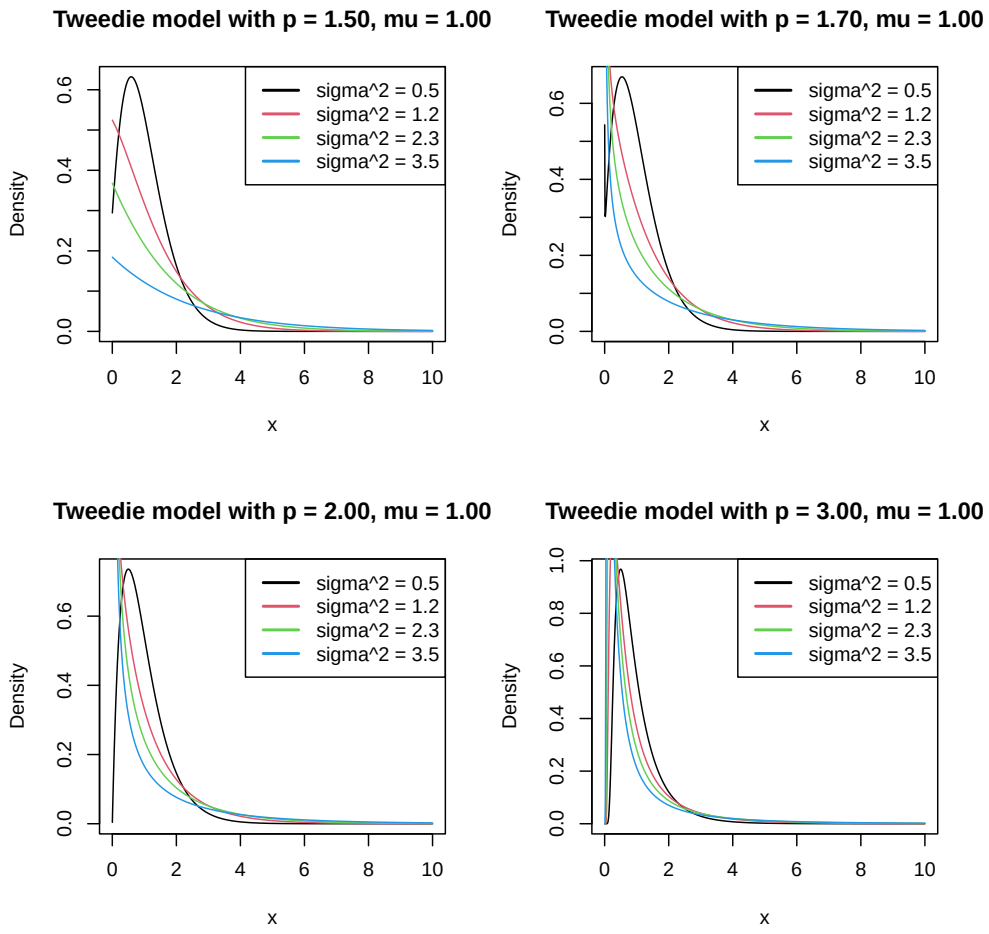

Fig. 2.1.2: The Tweedie distribution density plot with $\mu = 1$ and p, σ^2 varying

We know how to find the $\kappa(\theta)$ now, but it remains that the suitable functions $a(\cdot)$ and $b(\cdot)$ are still unknown, as we mentioned before that they can be any functions as long as they satisfy the Equation (2.1.1). However, the approximation of those functions can be evaluated by Fourier inversion of the characteristic functions, this idea is illustrated in details by Dunn and Smyth (2008). The only exceptions where $a(\cdot)$ and $b(\cdot)$ can be determined are when $p = 1$, $p = 2$ and $p = 3$ and especially at $y = 0$ when $1 < p < 2$. In this paper, we only concentrate on the case when $1 < p < 2$, as proposed by Dunn and Smyth (2008), the zero-mass value is given as:

$$f(0; \mu, \sigma^2) = \exp\left(-\frac{\mu^{2-p}}{\sigma^2(2-p)}\right). \tag{2.1.14}$$

Using Expression (2.1.14), we can draw a plot to see how zero-mass value correlated to the parameters in Tweedie models (see Figure 2.1.3). More intuitively, (Figure 2.1.4). gives the contour plot of Figure 2.1.3 which indicates that the higher the dispersion the higher the zero proportion and the higher the μ the lower the proportion.

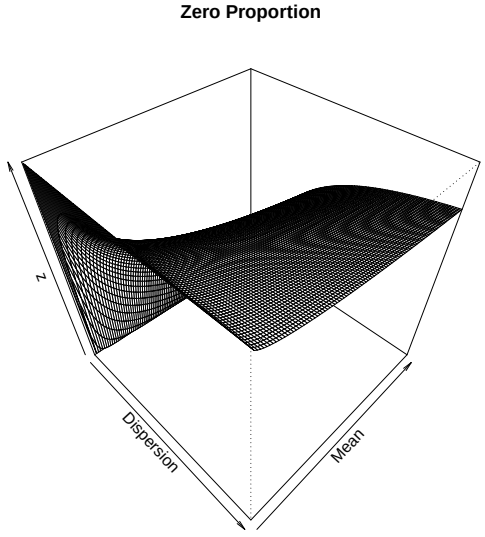

Fig. 2.1.3: Zero-mass values with varying $\sigma^2 \in [0, 100]$ and $\mu \in [0, 100]$

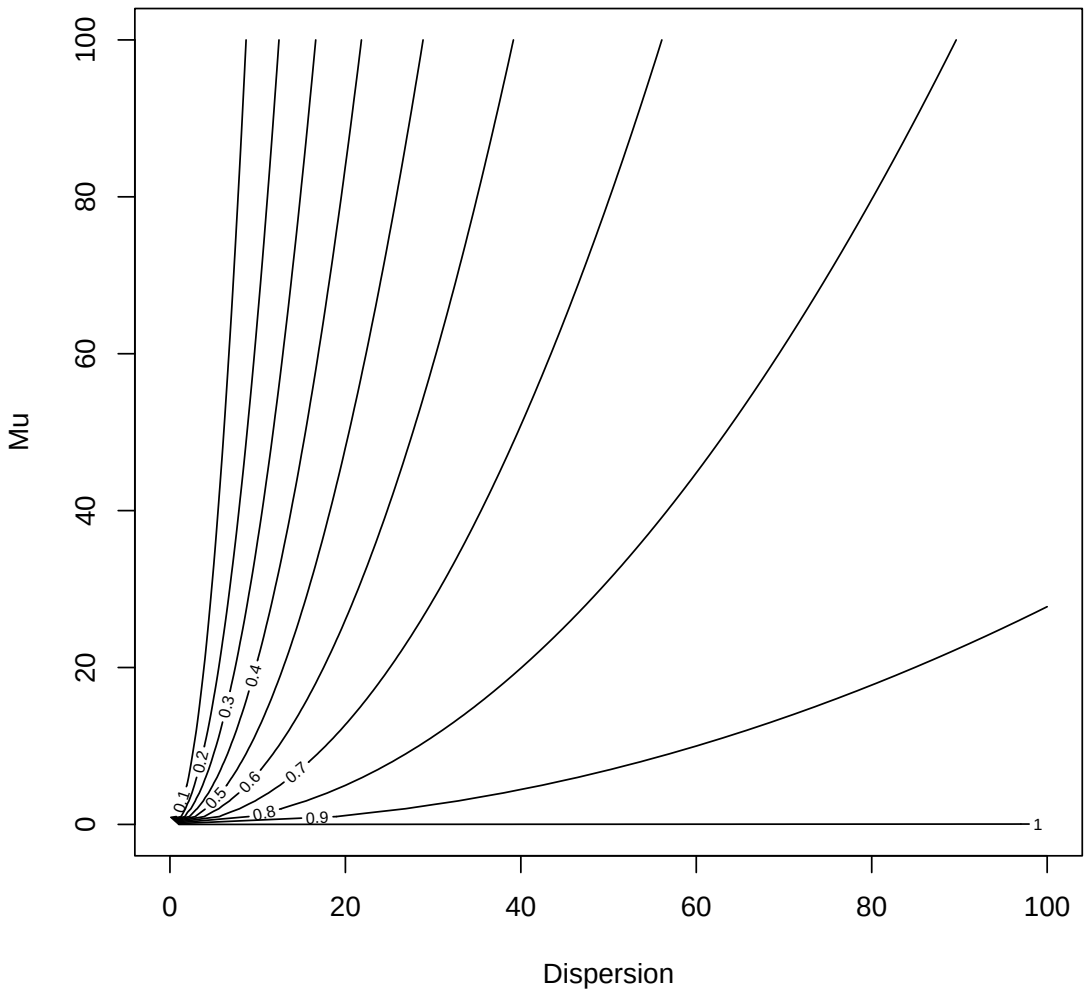

Fig. 2.1.4: Contour of the Figure 2.1.3

Dunn and Smyth (2008) demonstrate that the function mentioned in the paper exhibits strict convexity and can be approximated using Stirling's formula for the Gamma function and a Fourier inversion method for the infinite series. In practical applications, the parameters σ^2 and p are first estimated by maximizing the profile likelihood numerically, while profiling out the mean parameter μ based on a given value of σ^2. Next, the mean parameter is estimated using a Generalized Linear Model (GLM) with the previously estimated σ^2. It is important to note that the Tweedie distribution does not have a closed-form expression, and the profile likelihood needs

to be computed using numerical optimization methods.

2.2 GAM models

GAM (Generalized additive model), as defined in Hastie and Tibshirani (1987), is a type of generalized linear model where the linear predictor is composed of the sum of smoothing functions of the covariates. GAM extends the concept of linear regression by allowing for more flexible relationships between the dependent variable and the independent variables. It is particularly useful when dealing with non-linear relationships, interactions, and complex patterns in the data. The fundamental idea behind GAMs is to model the relationship between the response variable and each predictor as a sum of smooth functions, also known as smoothing splines or basis functions, which has the form looks like in Hastie and Tibshirani (1987):

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \tag{2.2.1}$$

where y is the response variable, x_i are the predictors, and $g(\cdot)$ is the link function relating the expected value of y to the predictor variables via this structure. In our study, the link function is defined as log link: $\log(E(Y))$. $f_i(x_i)$ are the smooth non-linear functions. These smooth functions are flexible and can capture a wide range of non-linear patterns. The individual smooth functions are then combined additively to form the overall model.

In a GAM, the dependent variable is typically assumed to follow a distribution from the exponential family, such as Gaussian (for continuous variables), binomial (for binary variables), or Poisson (for count variables). The predictors can be continuous, categorical, or a combination of both. In our study, we suppose that the dependent variable follows the Tweedie distribution and Gamma distribution. As we mentioned before, GAMs consists of smoothing functions and can be seen as a large GLM. Thus, when we are estimating the GAMs, we need to estimate simultaneously the smoothing functions and the parametric terms in the model. In our case, we use

cubic regression spline to estimate the smoother, and for computation convenience purpose, we use RMLE (Restricted Maximum Likelihood Estimator) to estimate the smoothing parameter λ. Figure 2.2.1 shows how the performance of cubic spline, where we can see that almost all the data points are interpolated by the cubic spline.

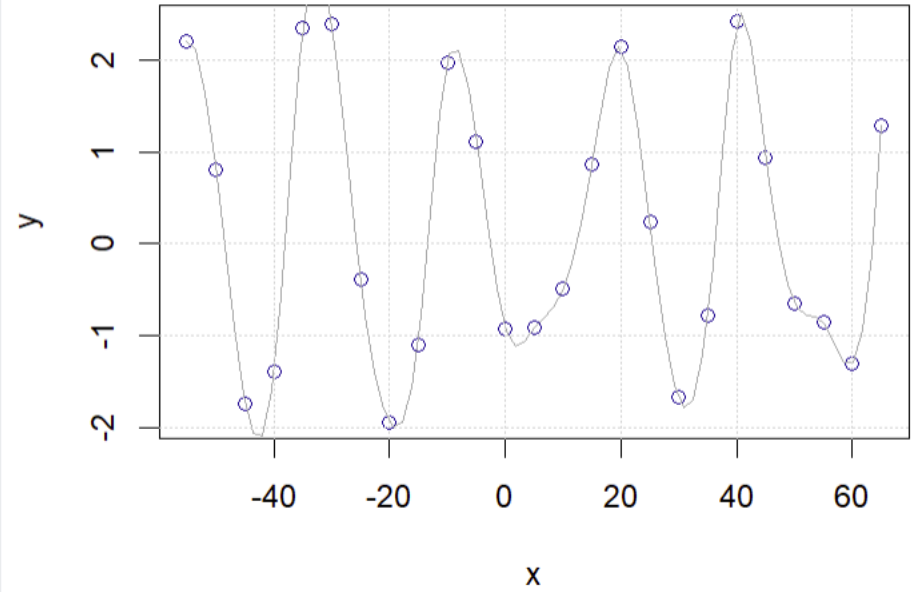

Fig. 2.2.1: Performance of cubic spline

In the paper by McKinley and Levine (1998), the process of cubic spline estimation in our case works as following:

Suppose we have data points x_i generated from the uniform distribution and y_i from gamma distribution $(x_1, y_1), ..., (x_n, y_n)$, and our spline is defined as

$$S(x) = \begin{cases} s_1(x) & \text{if } x_1 \leq x < x_2 \\ s_2(x) & \text{if } x_2 \leq x < x_3 \\ \dots & \\ s_{n-1}(x) & \text{if } x_{n-1} \leq x < x_n, \end{cases} \tag{2.2.2}$$

where each $s_i(x)$ is defined in the form:

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \tag{2.2.3}$$

for $i = 1, 2, ..., n - 1$. Observe the Expression (2.2.3), we can see that there are 4 parameters a_i, b_i, c_i and d_i, which means that we need 4 equations to find the expression for s_i, and to obtain these 4 equations we can use the 4 properties of the cubic spline:

1. The piecewise function $S(x)$ will interpolate all data points,
2. $S(x)$ is continuous on the interval $[x_1, x_n]$,
3. $S'(x)$ is continuous on the interval $[x_1, x_n]$,
4. $S''(x)$ is continuous on the interval $[x_1, x_n]$.

From property 1, we can get that:

$$S(x_i) = y_i, \quad for \quad i = 1, 2, 3, ..., n, \tag{2.2.4}$$

Plug Expression (2.2.2) into Expression (2.2.3), we can know that on each knots we produce that:

$$\begin{aligned} S(x_i) &= s_i(x_i) \\ &= a_i(x_i - x_i)^3 + b_i(x_i - x_i)^2 + c_i(x_i - x_i) + d_i \\ &= d_i = y_i. \end{aligned} \tag{2.2.5}$$

For $i = 1, 2, ..., n$. Since $S(x)$ is continuous across all the intervals at each knot, we must have:

$$\begin{aligned} s_{i-1}(x_i) &= s_i(x_i) \\ d_i = a_{i-1}(x_i - x_i)^3 + b_{i-1}(x_i - x_i)^2 + c_{i-1}(x_i - x_i) + d_{i-1}&. \end{aligned} \tag{2.2.6}$$

Letting the length of interval $h = x_{i-1} - x_i$ is equal for all the sub-intervals, $i = 2, 3, ..., n$, we can have:

$$d_i = a_{i-1}h^3 + b_{i-1}h^2 + c_{i-1}h + d_{i-1}. \tag{2.2.7}$$

Also, by property 3, we can have that the derivatives must be equal at the data points in order to make the curve smooth:

$$\begin{aligned} s_i'(x_i) &= s_{i-1}'(x_i) \qquad (2.2.8)\\ s_i'(x_i) &= 3a_i(x_i - x_i)^2 + 2b_i(x_i - x_i) + c_i = c_i \\ s_{i-1}'(x_i) &= 3a_{i-1}(x_i - x_{i-1})^2 + 2b_{i-1}(x_i - x_{i-1}) + c_{i-1} \\ \Rightarrow c_i &= 3a_{i-1}(x_i - x_{i-1})^2 + 2b_{i-1}(x_i - x_{i-1}) + c_{i-1} \\ c_i &= 3a_{i-1}h^2 + 2b_{i-1}h + c_{i-1}, \end{aligned}$$

for $i = 2, 3, ..., n-1$. By property 4, $s_i''(x)$ has to be continuous on all the intervals. Thus, we have that:

$$\begin{aligned} s_i''(x) &= 6a_i(x - x_i) + 2b_i \\ s_i''(x_i) &= 2b_i, \end{aligned}$$

for $i = 2, 3, ..., n-2$. Similarly as above, $s_i''(x_i) = s_{i-1}''(x_i)$ and we can have that:

$$\begin{aligned} s_{i-1}''(x_i) &= 6a_{i-1}(x_i - x_{i-1}) + 2b_{i-1} \\ \Rightarrow 2b_i &= 6a_ih + 2b_{i-1}. \qquad (2.2.9) \end{aligned}$$

Then, we can induce the expression for a_i as:

$$\begin{aligned} 2b_i = 6a_{i-1}h + 2b_{i-1} \Rightarrow 2b_i - 2b_{i-1} &= 6a_{i-1}h \\ \Rightarrow a_{i-1} &= \frac{b_i - b_{i-1}}{3h}, \qquad (2.2.10) \end{aligned}$$

and using Equation (2.2.7), the c_i can be rewritten as

$$\begin{aligned}
d_i &= a_{i-1}h^3 + b_{i-1}h^2 + c_{i-1}h + d_{i-1} \\
c_{i-1}h &= d_i - a_{i-1}h^3 - b_{i-1}h^2 - d_{i-1} \\
c_{i-1} &= \frac{d_i - d_{i-1} - a_{i-1}h^3 - b_{i-1}h^2}{h} \\
c_{i-1} &= \frac{d_i - d_{i-1}}{h} - (a_{i-1}h^2 + b_{i-1}h) \\
c_{i-1} &= \frac{d_i - d_{i-1}}{h} - (\frac{b_i - b_{i-1}}{3h}h^2 + \frac{s''_{i-1}(x_{i-1})}{2}h) \\
c_{i-1} &= \frac{d_i - d_{i-1}}{h} - (\frac{s''_i(x_i) - s''_{i-1}(x_{i-1})}{6h}h^2 + \frac{s''_{i-1}(x_{i-1})}{2}h) \\
c_{i-1} &= \frac{d_i - d_{i-1}}{h} - (\frac{s''_i(x_i) - s''_{i-1}(x_{i-1})}{6}h + \frac{3s''_{i-1}(x_{i-1})}{6}h) \\
c_{i-1} &= \frac{d_i - d_{i-1}}{h} - (\frac{s''_i(x_i) + 2s''_{i-1}(x_{i-1})}{6})h \\
c_{i-1} &= \frac{y_i - y_{i-1}}{h} - (\frac{s''_i(x_i) + 2s''_{i-1}(x_{i-1})}{6})h. \qquad (2.2.11)
\end{aligned}$$

We can now have enough equations to determine the weight for $n-1$ equations

$$\begin{aligned}
a_i &= \frac{s''_{i+1}(x_{i+1}) - s''_i(x_i)}{6h} \qquad (2.2.12) \\
b_i &= \frac{s''_i(x_i)}{2} \\
c_i &= \frac{y_{i+1} - y_i}{h} - \left(\frac{s''_{i+1}(x_{i+1}) + 2s''_i(x_i)}{6}\right)h \\
d_i &= y_i.
\end{aligned}$$

This system can be handled more easily by putting the formula into the matrix form, and the relative computation process can be found in McKinley and Levine (1998). Now, Expression (2.2.1) can be rewritten as

$$f_i(x) = \sum_{k=1}^{3} \beta_{ik} b_k(x), \qquad (2.2.13)$$

where $i = 1, 2, ..., p$, β_{ik} is the coefficient and $b_k(x)$ is the basis function, since in our case we are using cubic spline, $b_1(x)$ will be x, $b_2(x)$ is x^2 and $b_3(x)$ is x^3.

To estimate the GAM, we need to find the proper smoothing parameter λ, which is the parameter that controls the smoothness of the predictive functions. Generally we have 2 ways of estimating: Generalized cross validation criteria and RMLE. In our study, we will only use RMLE for it has less computation complexity. For both method, the core idea is to maximize the penalized likelihood function, the penalize likelihood function is given as in Wood (2004) by:

$$2\ell\left(\beta, f_1(x_1), f_2(x_2), ..., f_p(x_p)\right) - \text{penalty}. \tag{2.2.14}$$

The penalty can be expressed based on the second derivatives

$$\text{penalty} \ = \sum_{j=1}^{p} \lambda_j \int \left(s_j''(x_j)\right)^2 dx. \tag{2.2.15}$$

The parameter $\lambda_1, ..., \lambda_p$ are the smoothing parameters which controls how smooth our curve will be. Intuitively, we know that the second derivative measures the slope of the first derivative, that is to say how wiggle the curve will be, while larger second derivative will lead to a wiggler curve and a straight line will have 0 second derivative. Thus, it's reasonable to add up all the squared derivatives as the penalty, Figure 2.2.2 shows the impact of smoothing parameter directly.

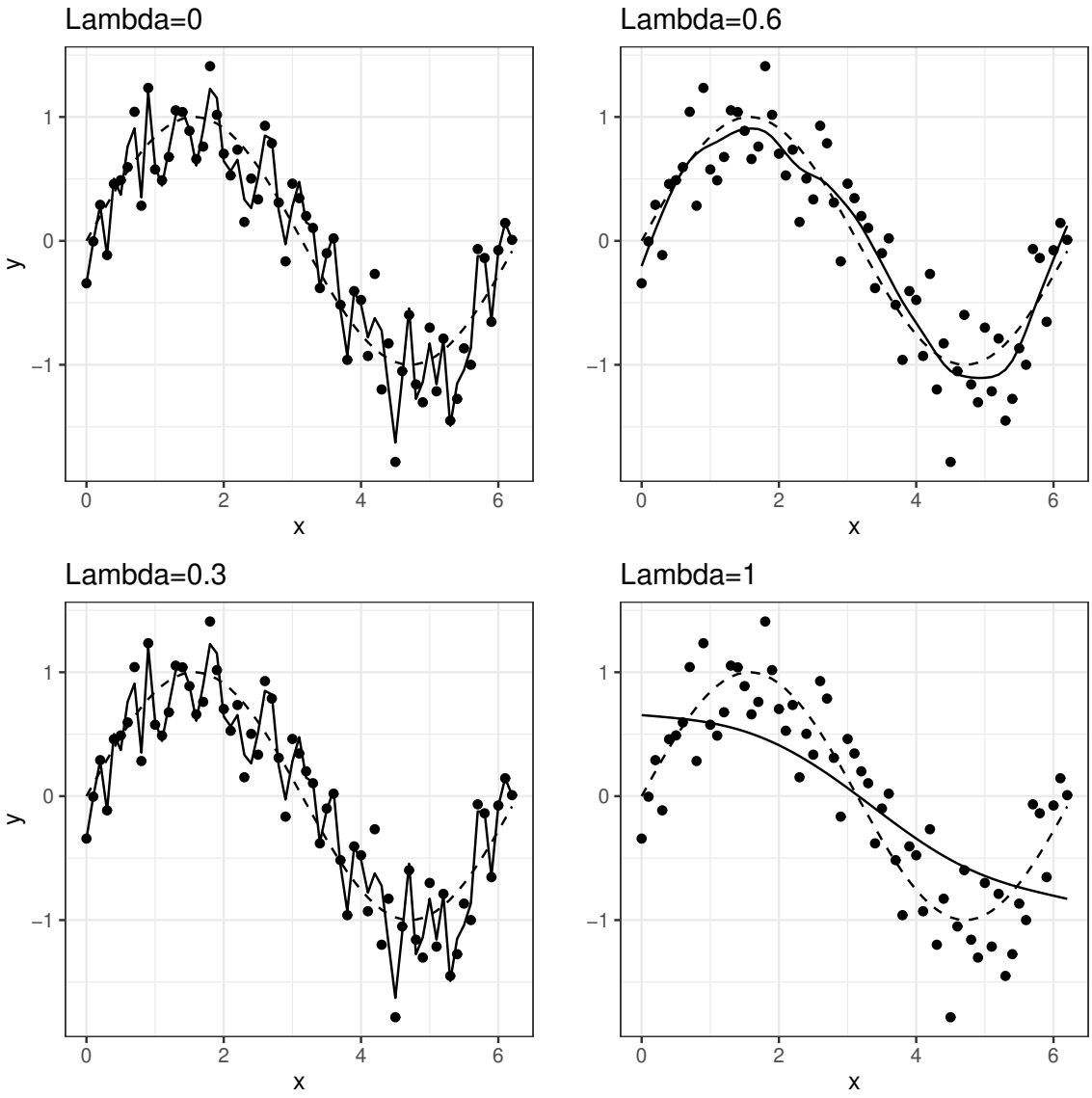

Fig. 2.2.2: Impact of smoothing parameter Larsen (2015)

To maximize the penalized likelihood function, we need to use RMLE, given the vector of smoothing parameter Λ, the restricted log likelihood function has the form Larsen (2015)

$$\ell_r\left(\hat{\beta}, \lambda\right) = \int f(y \mid \beta) f(\beta) d\beta, \tag{2.2.16}$$

which is taken by integrating out the joint pdf with respect to β, and the β is the matrix whose elements are coefficients in Expression (2.2.13). Firstly, for some given

random trail Λ, we estimate the β using penalized iterative re-weighted least square. Secondly, we update the Λ by maximizing the Expression (2.2.16). Repeat the above 2 steps until the result converges and then we obtain the estimation of smoothing parameters. More details about PIRLS (Penalized Iterative Re-weighted Least Square) method and how it works are stated in the paper of Wood (2004).

2.3 Two-part models

In our study, the response values are consisted of 2 parts, first part is $y = 0$ and the second part is $y > 0$. The outcome $y = 0$ is observed sufficiently frequent that those zeros can not be neglected. To address this problem, two-part model (2PM) assumes that the outcome of $Pr(Y = 0 \mid \mathbf{x})$ and $Pr(Y > 0 \mid \mathbf{x})$ is governed by a logistic regression. That is, suppose $Pr(Y = 0 \mid \mathbf{x}) = p$, then $\ln\left(\frac{p}{1-p}\right) = x_1\beta_1 + x_2\beta_2 +, ..., x_p\beta_p$. This is considered as the first part of modeling. Then for the part of $Pr(Y > 0 \mid \mathbf{x})$, we use GAM to fit the data and that is $\ln(E(Y \mid y > 0)) = \beta + f_1(x_1) + f_2(x_2) +, ..., + f_p(x_p)$. In this model, the outcome $\ln(y)$ is only observed when $y > 0$. The distribution of p is suppose to follow a Bernoulli distribution and those positive observations are supposed to from the Gamma distribution. In this paper, we are estimating the mean of the response variable, and the value in 2PM can be estimated as

$$\begin{aligned} E(Y) &= Pr(Y = 0 \mid \mathbf{x}) \times 0 + Pr(Y > 0 \mid \mathbf{x}) \times E(Y \mid y > 0) \\ \Rightarrow E(Y) &= Pr(Y > 0 \mid \mathbf{x}) \times E(Y \mid y > 0). \end{aligned} \tag{2.3.1}$$

More details of 2PM can be found in Mullahy (1998).

CHAPTER 3

Simulation study

In our study, we generate the data to simulate the GAM, the dependent variable Y is generated from the gamma distribution and the mean and θ parameter are defined as functions of the independent variables X_0 and X_1. Both independent variables are generated from the uniform (0,1) distribution, and the functions defined to generate the mean and θ are as following

$$\mu = \exp\left(\frac{2sin(\pi x_0)}{5}\right), \tag{3.0.1}$$

$$\theta = \exp\left(\frac{e^{x_1}}{2} - 2\right). \tag{3.0.2}$$

The dependent variable Y is generated from the $Gamma\left(\frac{1}{\theta}, \mu\theta\right)$, where the shape parameter is $\frac{1}{\theta}$ and the scale parameter is $\mu\theta$. We generate 1000 points by the process depicted above, Figure 3.0.1 and 3.0.2 show the relationship between X_1 and Y, X_0 and Y respectively. To make things more clear, we generate only 100 points to plot these 2 figures. From the plot, we can see that for both variables, they are not showing a linear relationship with Y. Thus, the linear models may not be very suitable here and GAM seems to be a good option in the case.

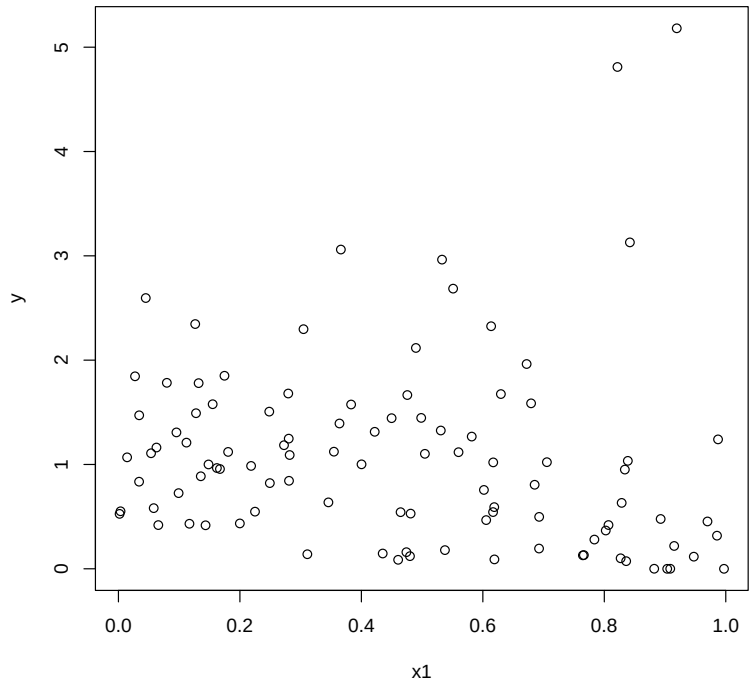

Fig. 3.0.1: Relationship between X_1 and Y

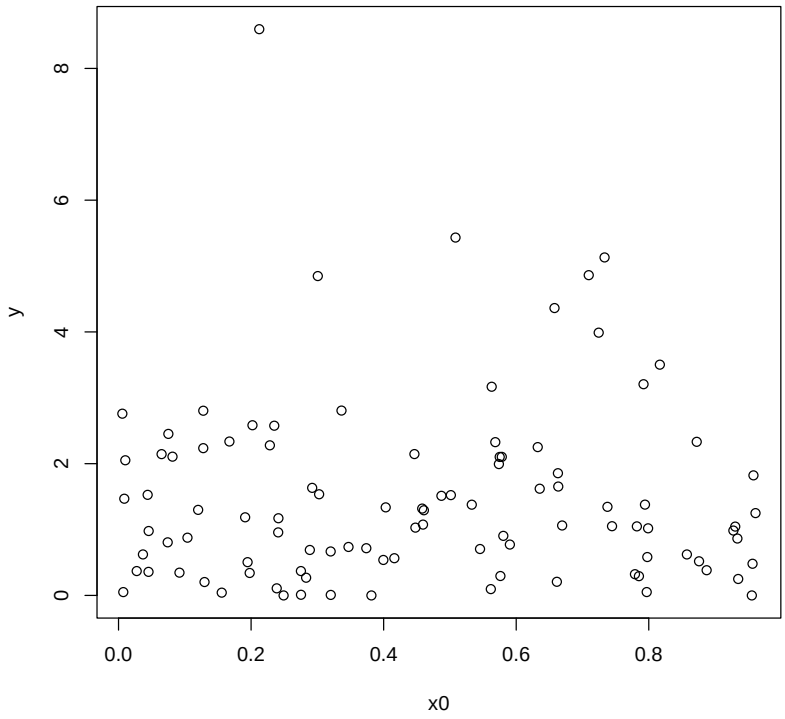

Fig. 3.0.2: Relationship between X_0 and Y

We assume that the zero proportion in our data has 4 different groups: 5%-15%, 15%-25%, 25%-35% and 35%-45%, the proportion of zero is represented by p. To create the zero observations, we generate a vector of N elements from the Bernoulli distribution and the mean of this vector is $1 - p$, then we multiply the observations y by this vector, thus we obtain the vector of responses which will contain zeros with probability p. Since later we will be estimating p with logistic regression, p is defined as a function of X_1, through the usual logit function:

$$p = \frac{1}{1 + \exp(\beta + \alpha x_1)}, \tag{3.0.3}$$

where β and α are taking values as (1.73, 1.27), (1.1, 0.65), (0.6,0.5) and (0.2, 0.4) respectively. Taking these values in pair guaranteed that the value of p is located properly in the range we set at the beginning. The corresponding distributions of p are shown in Figure 3.0.3 and Figure 3.0.4 shows the relationship between X_1 and p. It's clear that the values of p are located properly in the range we set before, and X_1 has a almost linear relationship with p.

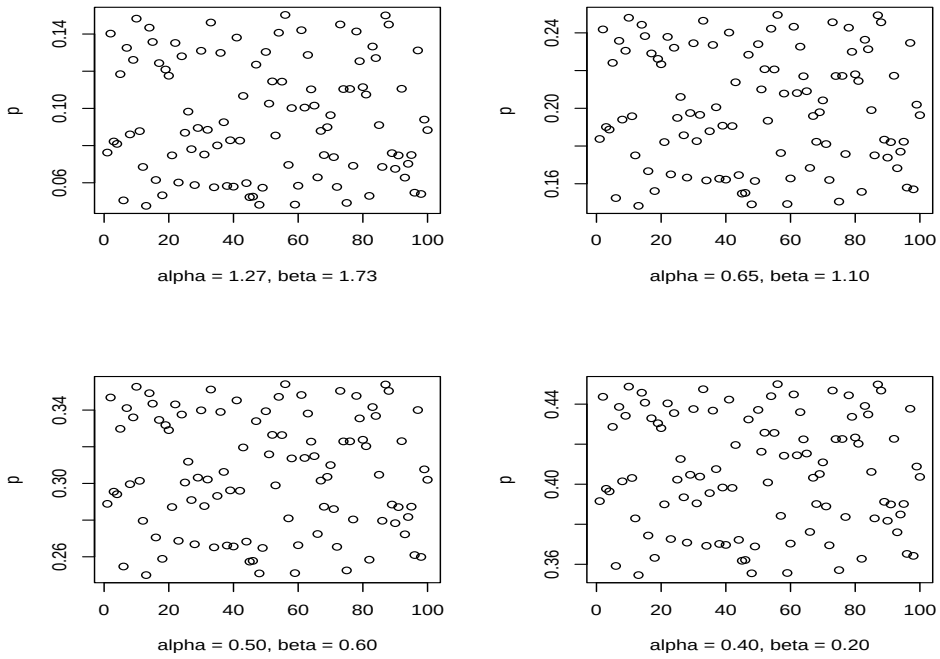

Fig. 3.0.3: Distribution of p

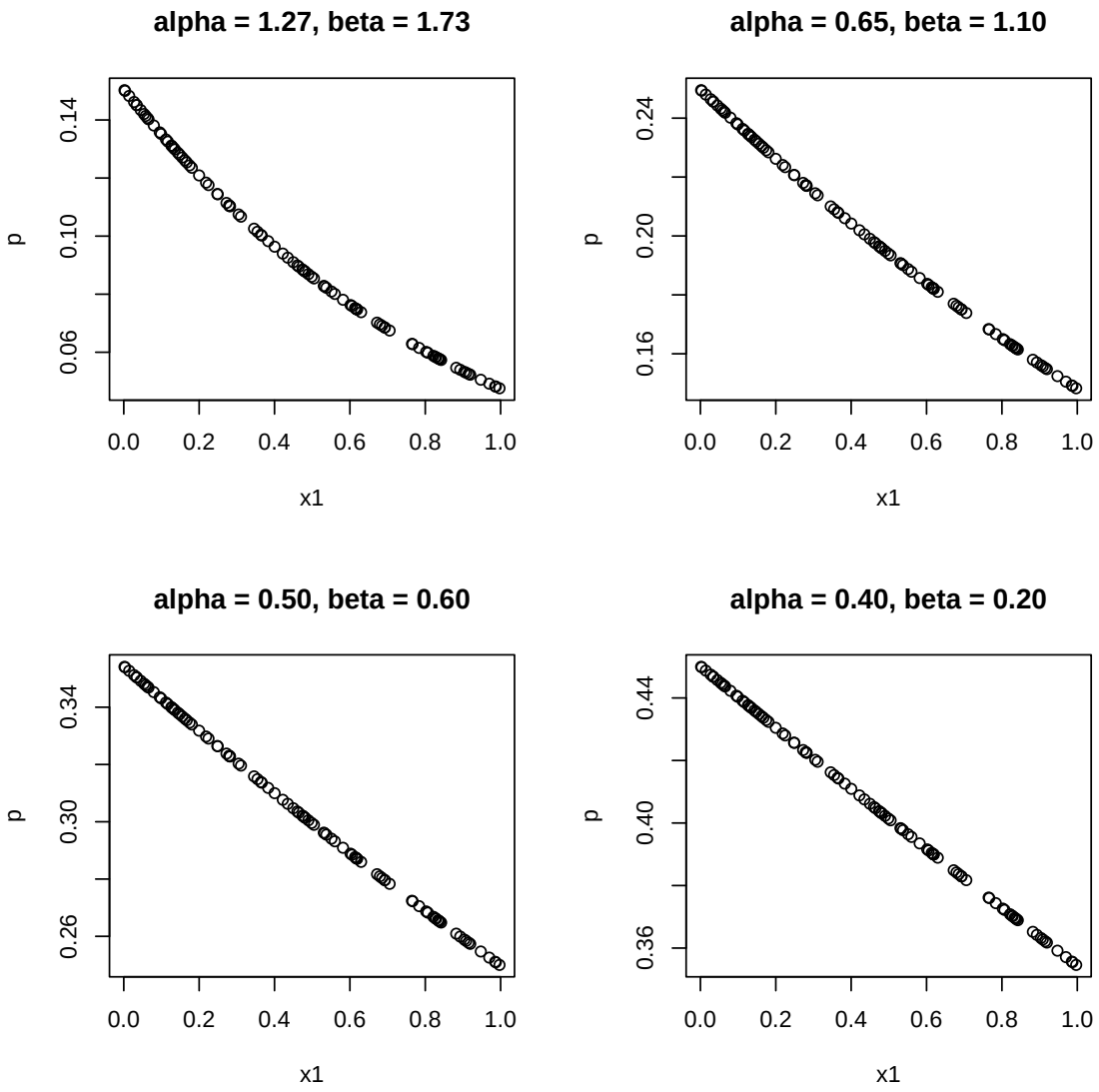

Fig. 3.0.4: X_1 and p

We use Monte Carlo simulation to address these 4 groups of data points, each group of data points are simulated 500 times. To evaluate the performance of both methods, we compare the mean predicted by using 2PM and simply Tweedie distribution. As comparison metric, we use MSE and mean integrated percent bias for predictive accuracy, the mean integrated percent bias in each run is defined as

follow:

$$Bias_j = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{\mu}_i - \mu_i|}{\mu_i}. \tag{3.0.4}$$

This expression looks pretty close to the formula for Mean Absolute Percentage Error (MAPE), despite we don't take the absolute value of the denominator term because the denominator in our case is strictly greater than 0, and note that it's different from the standard definition of bias. The MSE in each run is defined by

$$MSE_j = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_i - \mu_i\right)^2, \tag{3.0.5}$$

where $i = 1, 2, ..., N$, $j = 1, 2, ..., 500$ is the index for the running time of Monte Carlo and the $\hat{\mu}_i$ is the predicted mean by GAM and μ_i is the real mean value we generated in Expression (3.0.1). Since we are running the simulation 500 times, we take the average of these two metric. Thus, the overall metrics have the expressions

$$Bias = \frac{1}{500}\sum_{j=1}^{500} Bias_j, \tag{3.0.6}$$

$$MSE = \frac{1}{500}\sum_{j=1}^{500} MSE_j. \tag{3.0.7}$$

Also we set 3 different size data groups: $N = 500$, $N = 1000$ and $N = 2000$. We would like to see how these two methods work under each situation. The results of the simulation study are shown in the figures and tables below.

3.1 Presentation of the result

Table 3.1.1: The result of MSE and integrated percent bias using method of Tweedie distribution

N	5%-15%		15%-25%		25%-35%		35%-45%	
	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$N = 500$	0.0140	0.0808	0.0132	0.0887	0.0124	0.0996	0.0116	0.1126
$N = 1000$	0.0076	0.0594	0.0071	0.065	0.0074	0.0763	0.0059	0.0803
$N = 2000$	0.00507	0.0487	0.00469	0.0531	0.00435	0.0588	0.00406	0.0664

Table 3.1.2: The result of MSE and integrated percent bias using method of 2PM

N	5%-15%		15%-25%		25%-35%		35%-45%	
	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$N = 500$	0.0127	0.0767	0.0121	0.0849	0.0106	0.0920	0.0104	0.1063
$N = 1000$	0.0059	0.0525	0.0055	0.0577	0.0053	0.0649	0.0046	0.0704
$N = 2000$	0.00336	0.0395	0.00307	0.0425	0.00270	0.0465	0.00257	0.0528

From the above two tables, we can see clearly that as the number of data points increases, the figures of MSE and mean integrated percent bias decrease significantly and all the zero proportion groups are having the same trend.

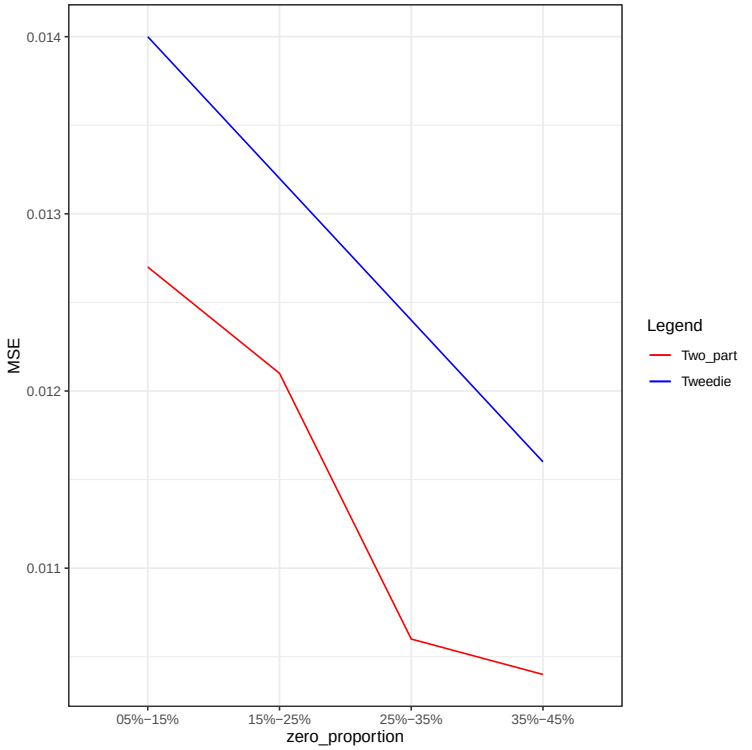

Fig. 3.1.1: MSE plot when $N = 500$

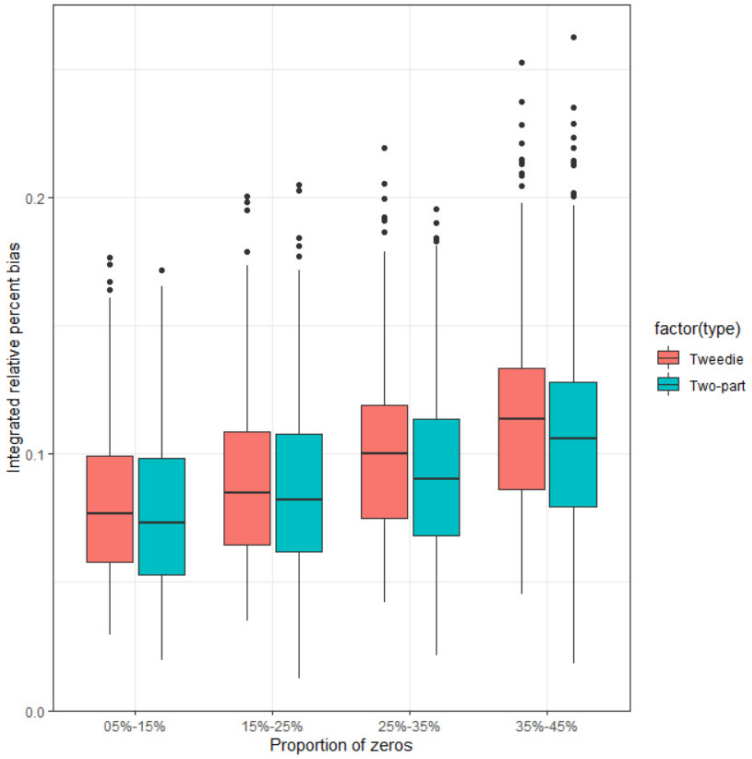

Fig. 3.1.2: Mean integrated percent bias plot when $N = 500$

When $N = 500$, we can see that the MSE of both methods are decreasing all the time as the zero proportion increases. Two-part model has smaller MSE than the model using Tweedie distribution. To see the distribution of mean integrated percent bias, we draw the boxplot of the integrated percent bias calculated by Expression (3.0.4) to see how are they distributed. From the Figure 3.1.2, we see that they both have an increasing trend and Two-part model has an overall smaller mean integrated percent bias.

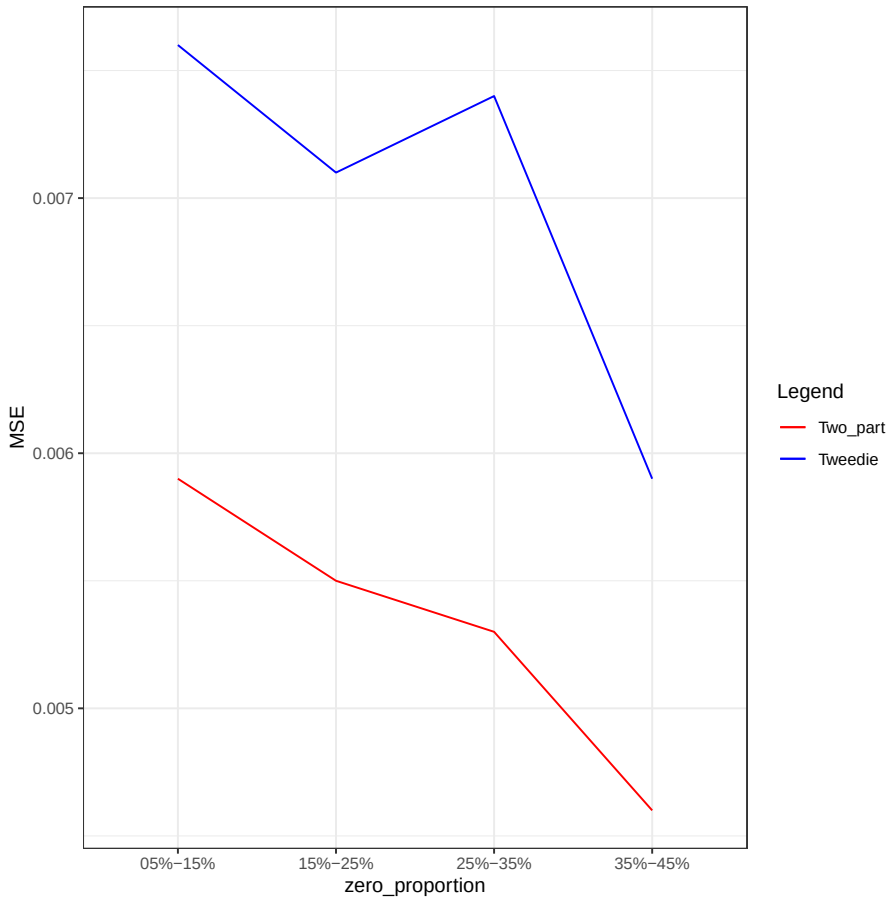

Fig. 3.1.3: MSE plot when $N = 1000$

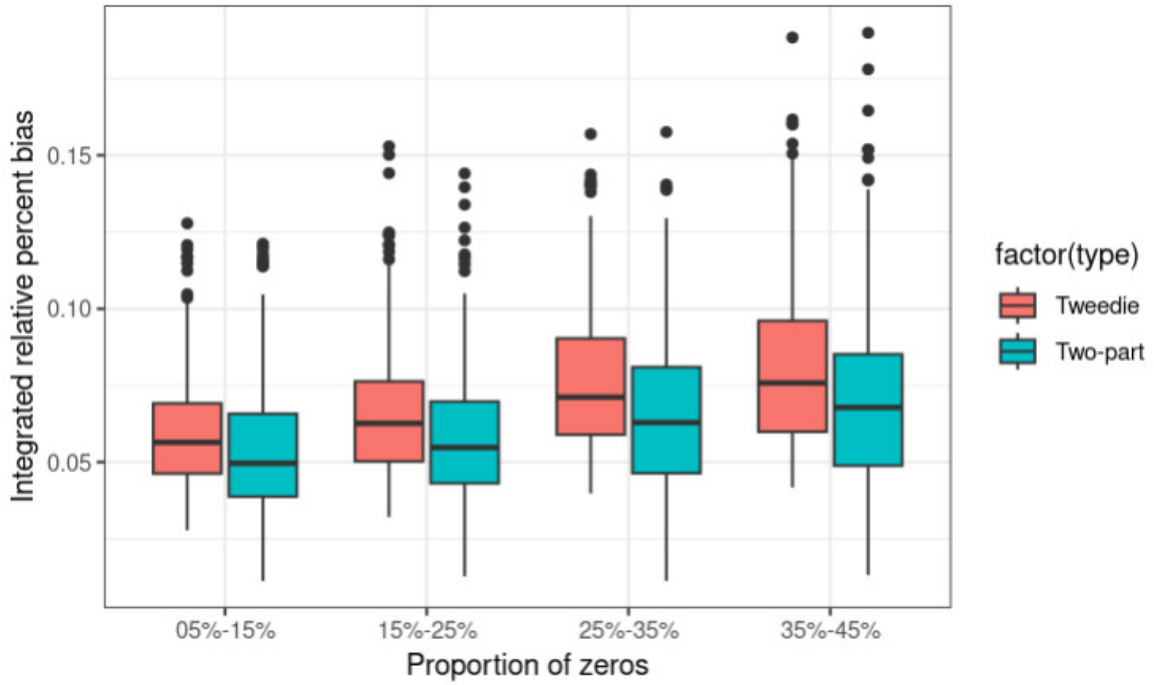

Fig. 3.1.4: Mean integrated percent bias plot when $N = 1000$

When $N = 1000$, the MSE plot of the Two-part model is decreasing as zero proportions increase. The MSE of the Tweedie distribution experienced a small increase when the proportion of zeros is 25%- 35% and then the curve goes down. The curve of the Two-part model is always below Tweedie. The variability of the mean integrated percent bias seems to be less for the case when $N = 1000$. For both methods, the percentage has an increasing trend, although Tweedie's method has always larger percentage.

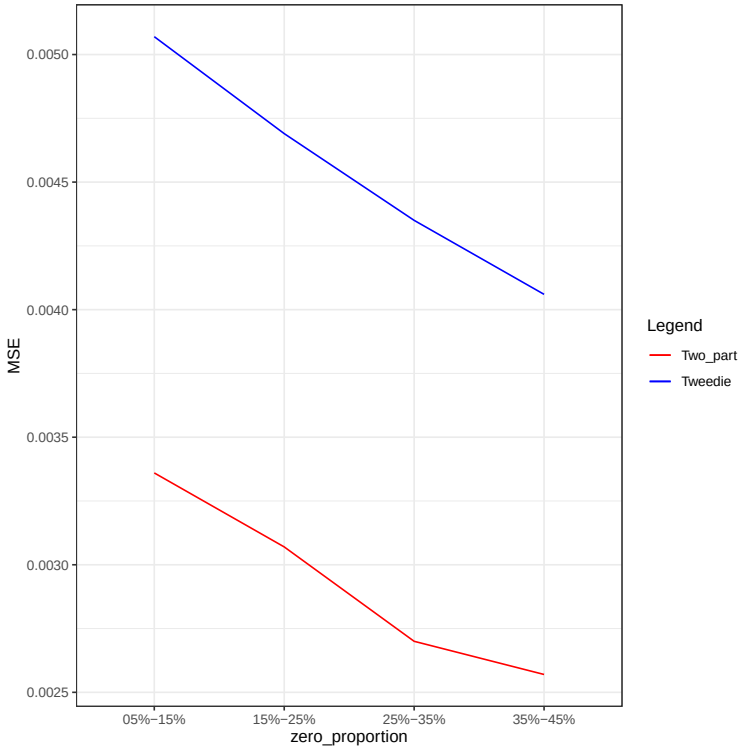

Fig. 3.1.5: MSE plot when $N = 2000$

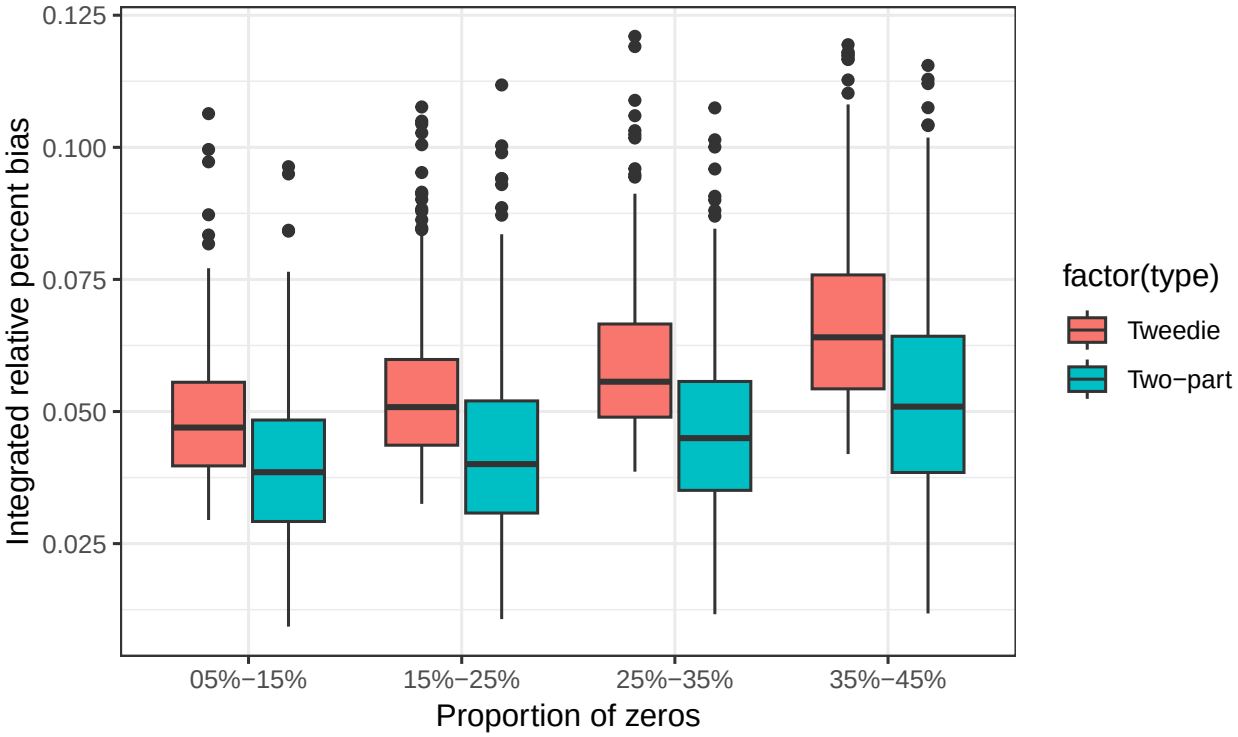

Fig. 3.1.6: Mean integrated percent bias plot when $N = 2000$

When $N = 2000$, the MSEs are decreasing for both models as the proportion of zeros increases. The Two-part model is always better in terms of MSEs. As for the mean integrated percent bias, it seems the variation does not change significantly compared to the case when $N = 1000$. The percentage increases for both methods as zeros increase, but the the Two-part performs better in all case.

CHAPTER 4

Conclusion and Future Work

The main objective of this paper is to compare the performance of the 2 methods for dealing with zero-inflated data. The first method is separating the fitting process into 2 parts, first estimate the probability of getting zero observations by logistic regression and then estimate the positive data points by GAM under the assumption of gamma distribution. Second method fits the data points by assuming a Tweedie distribution using GAM. The metrics we used to measure the performance of these 2 methods are mean relative bias and MSE computed from Monte Carlo simulations. From the results we obtained, we conclude that the Two-part model is a better option when dealing with positive, right-skewed data with excess zeros, since overall it has a smaller MSE and mean relative bias.

REFERENCES

Böhning, D., E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(2), 195–209.

Candy, S. (2004). Modelling catch and effort data using generalised linear models, the tweedie distribution, random vessel effects and random stratum-by-year effects. *Ccamlr Science 11*(0), 59–80.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse (1983). A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics 1*(2), 115–126.

Dunn, P. K. and G. K. Smyth (2008). Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing 18*(1), 73–86.

Foster, S. D. and M. V. Bravington (2013). A poisson–gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics 20*(4), 533–552.

Freund, D. A., T. J. Kniesner, and A. T. LoSasso (1999). Dealing with the common econometric problems of count data with excess zeros, endogenous treatment effects, and attrition bias. *Economics Letters 62*(1), 7–12.

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

Hastie, T. and R. Tibshirani (1987). Generalized additive models: some applications. *Journal of the American Statistical Association 82*(398), 371–386.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological) 49*(2), 127–162.

Jorgensen, B. (1997). The theory of dispersion models, chapman & hall. *CRC Monographs on Statistics and Applied Probability*.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Larsen, K. (2015). Gam: the predictive modeling silver bullet. *Multithreaded. Stitch Fix 30*, 1–27.

Liu, L., Y.-C. T. Shih, R. L. Strawderman, D. Zhang, B. A. Johnson, and H. Chai (2019). Statistical analysis of zero-inflated nonnegative continuous data. *Statistical Science 34*(2), 253–279.

McKinley, S. and M. Levine (1998). Cubic spline interpolation. *College of the Redwoods 45*(1), 1049–1060.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of health economics 17*(3), 247–281.

Raychaudhuri, S. (2008). Introduction to monte carlo simulation. In *2008 Winter simulation conference*, pp. 91–100. IEEE.

Shi, P. (2016). Insurance ratemaking using a copula-based multivariate tweedie model. *Scandinavian Actuarial Journal 2016*(3), 198–215.

Shono, H. (2008). Application of the tweedie distribution to zero-catch data in cpue analysis. *Fisheries Research 93*(1-2), 154–162.

Tu, W. (2006). Zero-inflated data. *Encyclopedia of environmetrics*.

Tweedie, M. C. et al. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, Volume 579, pp. 579–604.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*(467), 673–686.

VITA AUCTORIS

NAME:	Xianming Zeng
PLACE OF BIRTH:	China
EDUCATION:	Southern University of Science and Technology, B.Sc. degree in Statistics, 2017-2021
	University of Windsor, M.Sc in Mathematics and Statistics, Windsor, Ontario, 2023