Spring 5-22-2020

A Decision Tree Model to Predict Marginalized Zero-inflated Poisson Mean

Philip Amewudah
University of New Orleans, pamewuda@uno.edu

A Decision Tree Model to Predict Marginalized Zero-inflated Poisson Mean

A Thesis

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Master of Science
in
Mathematics

by

Philip Amewudah

M.S. University of New Orleans, 2020

May, 2020

Dedicated to Mum and Dad

Acknowledgements

I want to acknowledge the faculty members of the department of Mathematics, University of New Orleans, from whom I have learnt a lot, as far as Statistics is concerned. I particularly want to thank Dr. Xueyan Liu for her tremendous supervision and help to make this thesis a success.

Contents

List of Figures

List of Tables

List of Abbreviations

CART	**C**lassification **A**and **R**egression **T**ree
IDR	**I**ncedence **D**ensity **R**atio
ISLR	**I**ntroduction (to) **S**tatistical **L**earning (with) **R**
GSOEP	**G**erman **SO**cio**E**conomic **P**anel
MZIP	**M**arginalized **Z**ero-**i**nflated **P**oisson
ZIP	**Z**ero-**i**nflated **P**oisson

Abstract

Zero-inflated Poisson (ZIP) models are one of the most popular two-part models that are often employed to investigate the relationship between predictor variables and a count response that has more zeros than what is expected from the Poisson distribution. When inferences are targeted at marginal means, the ZIP models can be less effective, and interpretation of parameter estimates are easily misunderstood. Similarly, the first decision tree model that was developed to predict zero-inflated count response also does not also consider marginal effects of predictors on the entire population in its building process. In this research, we propose a marginalized ZIP (MZIP) decision tree model for predicting marginal means of overall population. Simulation studies were conducted to investigate the type 1 error, power and accuracy of the MZIP decision tree model. An application to real life data was demonstrated and results were provided to illustrate the applicability of our method.

Keywords: Decision tree, Marginalized Zero-inflated Poisson, Zero-inflated Poisson

1 Introduction

1.1 Zero-inflated Count Data

Count data arise from the outcome of count processes in continuous time. A classic example is the data collected as the number of incoming phone calls received from clients by a customer service personnel during a fixed time interval. One of the very common and widely known count process is the Poisson process. The underlying properties of this process are that, the outcomes of the events that make up the count data are independent of one another and that the probability of the occurrence of successive events are same. The distribution of the data collected from the Poisson process is known as the *Poisson distribution*. There are several other classical distributions of count data such as the Binomial distribution and the Negative Binomial distribution. In this thesis, the focus is the Poisson distribution. Some examples of data with Poisson distribution include, the monthly number of contracts strikes in U.S. manufacturing, (Kennan, 1985) and the number of patents of German companies registered at the German Patent Office in 1982, (Zimmermann and Schwalbach, 1991) etc.

The outcomes of count processes sometimes have excess zeros in real life applications. The data formed by these outcomes are said to be *zero-inflated*. Zero-inflated data sets are quite common in areas such as the environmental sciences and

the manufacturing applications, health services research and so on. When count datasets are zero-inflated, standard distributions usually are not able to accommodate the excess zeros. Considering the Poisson process for instance, we say the dataset is zero-inflated if there are more zeros than the Poisson distribution can model, where the zero-inflation leads to *overdispersion*, thus, variance greater than the mean. For example, when a manufacturing process is reliable, the count of defects on an item can be Poisson distributed. If the Poisson mean is λ, a large sample of n items should have about $ne^{-\lambda}$ items of no defects. However, sometimes there are many more items without defects than what would be predicted from the numbers of defects on imperfect items, (Lambert, 1992).

1.2 Modeling Zero-inflated Count Data

1.2.1 Statistical Models

The Poisson regression model has mostly been the benchmark traditional model for count data just as the normal linear model has been the benchmark for continuous data. When the count data is zero-inflated however, the Poisson regression model doesn't perform well because the modeling largely depends on the equidispersion assumption, which is commonly violated in real life. (Lambert, 1992) is the first to systematically develop the modeling of zero-inflated count data known as the *zero-inflated Poisson regression*. The ZIP model has two parts, one for the Poisson mean and the other latent part for the probability of being from a Poisson process. In the manufacturing example by (Lambert, 1992), the latent class effect is caused by unobserved changes that make the process randomly move back and forth between a perfect state that result in zero outcomes and an imperfect state that may or may not result in zero outcomes; thus the zero outcomes are from mixture of the perfect

states and the imperfect states. In contrast to the zero outcomes being a mixture of a perfect state and an imperfect state, (Mullahy, 1986) earlier described how hurdle models consider all zeros of the zero-inflated count data to be from the perfect state distinct from all non-zero outcomes.

There have been several other recent works on zero-inflated count data by other researchers. (Mullahy, 1997) showed that the unobserved heterogeneity assumed to be the source of over-dispersion in count data models has predictable implications for the probability structure of such mixture models. (Gupta, Gupta, and Tripathy, 2004) suggested score tests for a zero-inflated generalized Poisson model with applications to more diverse areas including the patent example, (Crepon and Duguet, 1997), road safety, (Miaou, 1994), species abundance, (Welsh et al., 1996), medical consultations, (Gurmu, 1997) and sexual behavior, (Heilbron, 1994).

Researchers sometimes would rather make inference about the marginal mean of the whole population than for two subpopulations with latent class interpretations as depicted by Lambert's ZIP model. Examples include population-based sample surveys aimed at describing an entire population, intervention studies that target populations where all members are considered to have some risk for the outcomes of interest or where interest is in the global effect in the population as a whole, (Long et al., 2014). (Albert, Wang, and Nelson, 2014) proposed estimators of overall effects using casual inference approaches related to the zero-inflated modeling framework. Literature on marginalized models indicate that analyst find the estimating of marginal effects of predictors different from the traditional ZIP model. (Heagerty, 1999) proposed marginalized multilevel models which model the marginal mean of the entire population by connecting marginal and conditional models with a function of covariates, marginal parameters and random effects specification. (Lee et al., 2011) investigated hurdle models in the context of marginalized

models to analyze clustered zero-inflated count data, marginalizing over the random effects. There have been several other papers that have explored marginal effects on an entire population other than separate effects of a two-part mixture model on subpopulations. (Liu et al., 2018) investigated the importance of goodness-of-fit evaluation and model selection in differentiating between the marginalized and non-marginalized models.

Motivated by the research in the literature, this thesis concentrates on the marginalized zero-inflated model first proposed by (Long et al., 2014) in which overall exposure effects estimates are easily obtained via a model for the marginal mean count.

1.2.2 Decision Tree Methods

The main area that this thesis focuses on is a non-parametric approach to modeling marginal mean of zero-inflated count data using the decision tree methods. Decision tree methods are very popular alternatives to parametric approaches and gaining more interest from practitioners due to its straightforward interpretation, less assumptions and high accuracy. First developed by (Breiman et al., 1984), the most popular decision tree methods are known as Classification and Regression Trees (CART) which employ recursive binary splitting of the predictor space and the splitting procedure can be demonstrated by a *dendrogram* that looks like a tree upside The tree grows from the top node, called root node to the bottom nodes called leaves. Nodes in the middle are called intermediate nodes. Branches connecting the nodes represent the splitting procedure. The method builds tree by a greedy search, that is, selecting the best splits from all possible splits at each intermediate node. At the terminal nodes, the average of the continuous response variable in the training set is used as the predicted value and the class with the larger proportion is often used as the predicted value for a categorical response variable.

When targeting count response variables, (Chaudhuri et al., 1995) proposed Poisson regression trees method which fits a log-linear model with the predictors at each intermediate node. The selection of the predictor for splitting is then based on a Levene's two sample test applied to each predictor to compare the positive and negative residuals. (Chaudhuri et al., 1995) also proposed generalized decision trees such as a logistic regression tree. (Loh and Shih, 1997) developed approaches to address unbiasedness of variable selection in the splitting process. For multivariate responses, (Segal, 1992) proposed a tree to handle continuous longitudinal responses. Several other contributions have been made that have not been covered in this review of the literature. Within the CART framework, (Therneau and Atkinson, 2004) developed the R package called `RPART`, which includes Poisson regression trees in R. The splitting criterion in based on the likelihood ratio test of the two Poisson groups from each potential split and the prediction of the response rate and the number of events at the node of the fitted tree.

Regarding zero-inflated count data, there is relatively little literature available. (Lee and Jin, 2006) proposed the ZIP tree models as a new decision tree tool to handle zero-inflated count data by using the ZIP log-likelihood as a homogeneity measure in the intermediate nodes. (Mathlouthi, Fredette, and Larocque, 2015) developed trees and random forests to predict ZIP responses which considered non-homogeneous Poisson processes.

1.3 Problem Statement and Objective of the Thesis

The ZIP tree model proposed by (Lee and Jin, 2006) produces confusing interpretations for predictions and effects of predictors that make splits because of the model structures. The effects of the whole population are not reflected in the results. As was the motivation of the traditional marginalized ZIP model, policy makers are

most of time interested in the marginal effects for decision making. The limitation in the ZIP tree model of (Lee and Jin, 2006) is the motivation for this thesis, given the relevance and advantages of tree models.

In this thesis, we aim to develop a decision tree method for marginalized zero-inflated distributions of zero-inflated count data which can bring straightforward interpretations about the effects of predictors to marginal means of the whole population.

1.4 Organization of Thesis

Chapter 2 reviews the Poisson, ZIP and the MZIP distributions and their log-likelihood functions. In Chapter 3, we go over the tree method for count data in the literature. Chapter 4 demonstrates the methodology of our MZIP tree models. Simulation results and empirical study are provided Chapter 5 and Chapter 6 respectively to illustrate the validity and applicability of our method. Conclusion and future work are discussed in Chapter 7.

2 Marginalized Zero-inflated Poisson Distribution

2.1 Introduction

In this chapter, we review the MZIP distribution and its log-likelihood function which will be used to build the decision tree in Chapter 3. Since the zero-inflated count data is a mixture of the Poisson and the Bernoulli distributions, we start off by going over these two distributions. Afterwards, the statistical ZIP model and the MZIP model will be reviewed.

2.2 The Poisson Distribution

If a count variable, Y, has a Poisson distribution denoted by $Y \sim \text{Poisson}(\lambda)$, where λ is the rate of occurrence within a particular time period, then the probability mass function, pmf, $f(y;\lambda)$, of the variable is given as

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, 3, ... \tag{2.1}$$

The Poisson distribution belongs to the family of distribution known as the exponential family of distributions. Generally, if a variable Y belongs to the exponential

family of distribution, then the distribution function of the variable is given as

$$f(y;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) , \tag{2.2}$$

where ϕ is the dispersion parameter and θ is the natural/canonical parameter with $\mathrm{E}(y) = b'(\theta)$ and $\mathrm{Var}(y) = b''(\theta)$. The canonical parameter becomes the link function that determines the relationship that a set of predictors can have with a response variable Y within the Generalized Linear Model (GLM) framework. For the Poisson distribution, $\theta = \log\ \lambda$, $a(\phi) = 1$, $b(\theta) = e^{\theta}$ and $c(y,\phi) = -\log(y!)$.

$$\begin{aligned}
\mathrm{E}(y) &= b'(\theta) = e^{\theta} = \lambda \, , \\
\mathrm{Var}(y) &= b''(\theta)a(\theta) = e^{\theta} = \lambda \, .
\end{aligned}$$

2.3 The Bernoulli Distribution

The Bernoulli distribution is another member in the exponential family of distributions. If the variable, D, has a Bernoulli distribution denoted by $D \sim \mathrm{Binomial}(1,p)$, then the pmf, $f(d;p)$, of the variable is given as

$$f(d;p) = p^{d}(1-p)^{1-d} \; ; \;\; d = 0,1 \, . \tag{2.3}$$

Compared to (2.2), the (2.3) can be re-written such that $\theta = \log\ \frac{p}{1-p}$, $a(\phi) = 1$, $b(\theta) = \log(1-p)$ and $c(y,\phi) = 0$. Therefore

$$\begin{aligned}
\mathrm{E}(y) &= b'(\theta) = p \, , \\
\mathrm{Var}(y) &= b''(\theta)a(\theta) = p(1-p) \, .
\end{aligned}$$

2.4 The Statistical Poisson and ZIP Regression Models

The Poisson regression model has been the benchmark model for data with count responses which assumes the equidispersion, thus, the mean count equals the variance in the counts. The Poisson regression stems from the GLM framework for modeling a response variable in the exponential family of distributions. In general, GLM uses a link function to provide the relationship between the linear predictors X_j, $(j = 1, 2, ..., p)$ and the conditional mean of the density function:

$$g[E(y|x)] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \,, \tag{2.4}$$

where $\beta_0, \beta_1, ...\beta_p$ are parameters and $g(\cdot)$ is the link function.

When Y_i's, $(i = 1, 2, ..., n)$, are independent and identically distributed (iid) and follow the Poisson distribution conditional on X_i's, we use $\log(\cdot)$ as the link function and call the model a Poisson regression model:

$$\log(\lambda_i|x_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \,. \tag{2.5}$$

When a count data has excess zeros, there is overdispersion and hence the equidispersion assumption of the Poisson is violated. The Poisson model is not an appropriate model for this situation anymore. (Lambert, 1992) proposed ZIP models that address the mixture of excess zeros and Poisson count process. The mixture is indicated by the latent binary variable d_i using a logit model and the density for the Poisson count given by the log-linear model. Thus,

$$y_i = \begin{cases} 0, & \text{when } d_i = 0 \,, \\ y_i^*, & \text{when } d_i = 1 \,, \end{cases}$$

where the latent indicator $d_i \sim$ Bernoulli(p_i) with $p_i = P(d_i = 1)$ and $y_i^* \sim$ Poisson(λ_i). The mixture yields the marginal probability mass function of the observed y_i given as:

$$f(y_i) = \begin{cases} (1-p_i) + p_i e^{-\lambda_i} & \text{if } y_i = 0 \\ p_i e^{-\lambda_i} \lambda_i^{y_i} / y_i! & \text{if } y_i = 1, 2, \dots, \end{cases}$$

where λ_i and p_i are modeld by

$$\begin{aligned} \log(\frac{p_i}{1-p_i}|Z_i) &= \gamma_0 + \gamma_1 z_{i1} + ... + \gamma_q z_{iq} = \boldsymbol{z_i^T \gamma}\,, \\ \log(\lambda_i|X_i) &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \boldsymbol{x_i^T \beta}\,, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\gamma}$= $(\gamma_0, \gamma_1, ...\gamma_q)^T$ is a vector of parameters and $\boldsymbol{z_i} = (1, z_{i1}, z_{i2}, ..., z_{iq})^T$ are the realizations of covariates in the logit model for p_i and $\boldsymbol{\beta}$= $(\beta_0, \beta_1, ...\beta_p)^T$ is the parameter vector associated with the covariates $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, ..., x_{ip})^T$ in the log-linear model for λ_i, $i = 1, 2, ..., n$. The likelihood function for this ZIP model is given by

$$L(\boldsymbol{\beta, \gamma}|y_i, x_i, z_i) = \prod_{y_i=0} [1 - p_i + e^{-\lambda_i} p_i] \cdot \prod_{y_i>0} [p_i \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}]\,, \tag{2.7}$$

where $\lambda_i = e^{\boldsymbol{x_i^T \beta}}$ and $p_i = e^{\boldsymbol{z_i^T \gamma}} / (1 + e^{\boldsymbol{z_i^T \gamma}})$.

(Long et al., 2014) pointed that the marginal mean $v_i = E(y_i|x_i, z_i) = p_i \lambda_i$ and hence

$$v_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{-(\gamma_0 + \gamma_1 z_{i1} + ... + \gamma_p z_{ip})}}\,. \tag{2.8}$$

According to (Long et al., 2014), unless $\gamma_j = 0$, for $j = 1, 2, ..., q$, the incidence density ratio (IDR), i.e., the ratio of marginal mean for a one unit increase in the j^{th} predictor, x_{ij}, will not be constant across the various level of the covariates in the logistic portion of the ZIP model. The fact can be seen from the following equation,

where we let $x_i = z_i$:

$$\frac{\mathrm{E}(y_i|x_{ij} = l+1)}{\mathrm{E}(y_i|x_{ij} = l)} = e^{\beta_j} \frac{1 + e^{-\boldsymbol{x_i}^T \boldsymbol{\gamma}}}{1 + e^{-(\gamma_j + \boldsymbol{x_i}^T \boldsymbol{\gamma})}} , \tag{2.9}$$

This challenge motivated (Long et al., 2014) to introduce the marginalized ZIP model.

2.5 The Marginalized ZIP (MZIP) Model

(Long et al., 2014) admitted the mixture of structural zeros and the Poisson count in ZIP and used the same model for the latent participation indicator but the log-linear model for the marginal mean v_i instead of the Poisson mean λ_i:

$$\begin{aligned} \log(\frac{p_i}{1-p_i}|z_i) &= \gamma_0 + \gamma_1 z_{i1} + ... + \gamma_p z_{iq} , \\ \log(v_i|x_i) &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} . \end{aligned} \tag{2.10}$$

They called it the MZIP model. It implies immediately that, e^{β_j} is the amount by which the overall mean, v_i, is multiplied for a unit change in X_j. Similar to (2.8) the Poisson density is given by

$$\lambda_i = \frac{v_i}{p_i} = e^{\boldsymbol{x_i^T \beta}} (1 + e^{-\boldsymbol{z_i^T \gamma}}) . \tag{2.11}$$

The likelihood function of the MZIP model can be derived from (2.7) by substituting $p_i = e^{z_i^T\gamma}/(1+e^{z_i^T\gamma})$ and $\lambda_i = e^{x_i^T\beta}(1+e^{z_i^T\gamma})$ to give the expression:

$$\begin{aligned} L(\gamma, \beta | y_i, x_i, z_i) = &\prod_{\text{all i}} (1+e^{z_i^T\gamma})^{-1} \prod_{y_i=0} (1+e^{z_i^T\gamma - (1+e^{-z_i^T\gamma})e^{x_i^T\beta}}) \\ &\prod_{y_i>0} [e^{z_i^T\gamma - (1+e^{-z_i^T\gamma})e^{x_i^T\beta}} (1+e^{-z_i^T\gamma})^{y_i} e^{x_i^T\beta y_i}/(y_i!)] \,. \end{aligned} \tag{2.12}$$

So far in this chapter, we have introduced the traditional Poisson related statistical models based on which the decision tree models would be built. In keeping with the aim of this thesis, decision tree models would be introduced, and their methodologies will be explained in the next chapter.

3 Decision Tree Models

3.1 Introduction

In this chapter, decision tree models will be studied. We would look at the fundamental algorithm of the Classification And Regression trees (CART) and how the CART splitting criteria work. Also, we will review how the Likelihood ratio test is adopted into the splitting criteria for decision trees and extended to accommodate ZIP Poisson data.

3.2 Classification And Regression Trees (CART)

Decision trees are one of the common data mining tools for classification and prediction. They have the advantage of being very easy to interpret and very flexible at modeling data. In this thesis, we develop our binary decision tree models based on the CART algorithm due to (Breiman et al., 1984). In this section, we use the regression trees to explain the algorithm of CART.

Consider for instance the regression problem where a baseball player's annual salary in thousands of dollars is predicted based on the number of years that he has played, `Years`, and the number of hits he has made in the previous year `Hits`. This example can be found in the Introduction to Statistical learning with R (ISLR) book,

(James et al., 2013) and the dataset, `Hitters`, is in the `ISLR` package in R. A decision tree model for this data is shown in Fig. 3.1. The tree consists of a series of binary

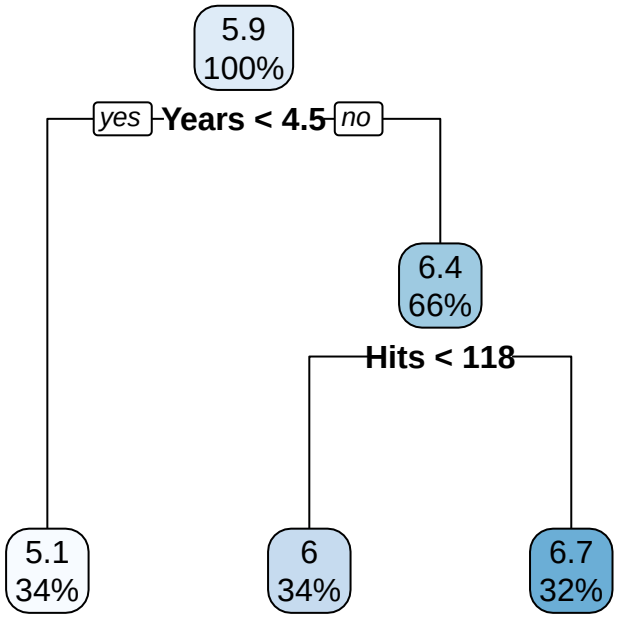

FIGURE 3.1: *For the* `Hitters` *data, a regression tree for predicting the logarithm of a baseball player's annual salary, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. The first split at the top of the tree results in two large branches. The left-hand branch corresponds to* `Years` < 4.5*, and the right-hand branch corresponds to* `Years` ≥ 4.5*. The tree has two intermediate nodes and three terminal nodes, or leaves. The number shown in each leaf is the mean of the response of the training set that fall in that region.*

splitting rules to divide the predictor space into disjoint subregions. The top split assigns observations having *Years* < 4.5 to the left child node. The predicted salary for the players who have played baseball in major leagues in less than 45 years is given by the mean response value of the players in the training data with `Years` $<$ 4.5. For such players, the mean logarithm of their annual salary in thousands of dollars is 5.1, and so we make a prediction of $e^{5.1} \approx 164$ thousands of dollars, for the player that belongs to that group. Those players who have `Years` ≥ 4.5 are

assigned to the right node, and then that group is further divided by the number of hits the players made in the previous year `Hits`. In all, this decision tree stratifies the players into three disjoint regions of predictor space: players who have played for fewer than four and a half years (R_1), players who have played for five or more years and made less than 118 hits last year (R_2), and players who have played for five or more years and made at least 118 hits last year (R_3). The predicted salaries for these three groups are $\$1,000 \times e^{5.1} = \$164,022$, $\$1,000 \times e^{6} = \$403,428$ and $\$1,000 \times e^{6.7} = \$812,406$ respectively.

The regions $R_1, R_2,$ and R_3 are known as the terminal nodes or leaves of the tree. The nodes along the tree where the predictor space is split are referred to as internal nodes. From Fig. 3.1, the one internal node is indicated by `Hits` < 118. The segments of the tree that connect the nodes are referred to as the branches.

In terms of interpretation, Fig. 3.1 implies that, `Years` is the most important factor in determining a player's annual salary, and players with little experience earn lower salaries than more experienced players. Given that a player has little experience, the number of hits that he made in the previous year play little role in his salary. But among players who have played for five or more years, the number of hits made in the previous year does affect salary, and players who made more hits last year tend to have higher salaries.

From the above example, we can understand that the CART algorithm involves stratifying and segmenting predictor spaces into a number of disjoint and rectangular regions/terminal nodes and using the mean of the observed responses in the training set in that region as the prediction for new data points that fall in that region. The algorithm consists of basically two steps:

1. split the predictor space in a binary fashion using the set of possible values for predictors $X_1, X_2, ..., X_p$, into J distinct and non-overlapping regions,

$R_1, R_2, ..., R_J$, thus, J terminal nodes.

2. for every observation that falls into the region R_j , make the prediction, the mean of the responses in that R_j in the training data.

The goal is to find the regions $R_1, ..., R_J$ that minimize the Residual Sum of Squares, (RSS), the cost function given by

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 , \tag{3.1}$$

where $\hat{y}_{R_j}$ is the prediction for region R_j: $\hat{y}_{R_j} = \text{Ave}_{i \in R_j} y_i$.

3.2.1 Recursive Binary Splitting

It is computationally challenging to consider every possible partition of the predictor spaces in all J regions to get the global minimum of RSS; as a result, the recursive binary splitting using a greedy search is utilized from the top of the tree to the bottom. To perform recursive binary splitting, a predictor X_j and a cutpoint s is also selected such that splitting the predictor space into the regions $R_1 = \{X | X_j < s\}$, the region of predictor space in which X_j takes on a value less than s and $R_2 = \{X | X_j \geq s\}$, the region of predictor space in which X_j takes on a value greater than or equal to s, leads to the greatest possible reduction of RSS. The selection is based on the consideration of all predictors and every possible cutpoint. Hence the values j and s are sought such that the RSS

$$\sum_{i: x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2} (y_i - \hat{y}_{R_2})^2 , \tag{3.2}$$

is minimized, where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the mean responses of the observations in R_1 and R_2 respectively. A node with descendant nodes is known a parent or internal node and binary split at the node yields are known as leftchild node and right child node. This binary splitting is repeated until the predictors and cutpoints are selected that minimizes RSS in all regions.

3.2.2 Tree Prunning

The tree example in Fig. 3.1 is actually a smaller tree known as a pruned tree from a bigger tree with more terminal nodes. The recursive binary splitting explained above naturally produces a very big tree that overfits the training dataset and hence performs poorly on test datasets in terms of prediction accuracy. A smaller tree with fewer terminal nodes is preferred such that the variance in the prediction is reduced to improve performance on test datasets at the expense of an increase in bias. This process of deriving a smaller tree from the bigger tree is known as pruning. One way of pruning a decision tree is by introducing a tuning parameter known as the complexity parameter. For a complexity parameter denoted by cp, a subtree $T \subset T_0$ is obtained such that

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + cp|T| \tag{3.3}$$

is small as possible where $|T|$ is the number of terminal nodes of the pruned subtree T, R_m is the region corresponding to the m^th terminal node, and $\hat{y}_{R_m}$ is the predicted response associated with R_m. The complexity parameter in this case adds a penalty to the cost function using the tree size and hence controls the complexity of the pruned tree and its fit to the training data. When cp $= 0$, there is no pruning at all. As cp increases from 0, the branches of the big tree get pruned. The choice of the most appropriate value of cp can be made through k-fold cross-validation.

3.3 Goodness of Split Criteria

From the CART algorithm, the major technique of the binary recursive splitting explained so far is to develop a tree that reduces the RSS or the overall variance in prediction as much as possible. In the tree building process, equivalent to the variance reduction technique in binary recursive splitting is the maximization of a goodness of split measure defined by the difference of the mean square error between the parent node and the child nodes as shown in (Breiman et al., 1984). For most statistical software like `R`, this goodness of split measure is the underlying measure that is used to perform the binary recursive splitting of the decision tree model. For the regression tree, this goodness of split measure, ϕ, focuses on variance reduction defined as

$$\phi_{\text{CART}} = V_p - \frac{n_l}{n} V_l - \frac{n_r}{n} V_r \,, \tag{3.4}$$

where $V_{R_p} = 1/n_p \sum_{i:x_i \in R_p} (y_i - \hat{y}_{R_p})^2$, is the variance of the responses in the parent node; $V_l = 1/n_l \sum_{i:x_i \in R_l} (y_i - \hat{y}_{R_l})^2$ is the variance of the responses in the left child node and $V_r = 1/n_r \sum_{i:x_i \in R_r} (y_i - \hat{y}_{R_r})^2$ is the variance of the responses in the right child node; n_p, n_l and n_r are the number of observations in the parent, left and right nodes respectively; R_p, R_l and R_r represent the parent, left and right nodes/regions. The split that results in a maximum value of the goodness of split measure is selected as indicated earlier.

3.3.1 Deviance

An important observation of the variance reduction technique of the CART algorithm explained above for the regression tree is that, the goodness of split measure is an expression of the mean square error about the mean in the various nodes in

the splitting process. Equivalently, the sum of square errors about the mean of the responses can also be used in place of the mean square error. This sum of squares error is actually the deviance of a regression model that is fit to predict a normally distributed response variable. (Therneau and Atkinson, 2019) provided a general expression of a goodness of split measure using the difference between the within-node deviance of the response data in the parent group, D_p and the sums of the within-node deviance of the response data in the left and right child group, D_l and D_r given as

$$\phi = D_p - D_l - D_r \, . \tag{3.5}$$

The split that maximizes (3.5) is sought. Since this is a generic case, the within deviance is specified based on the underlying distribution of the response variable.

3.4 Poisson trees

Having described the algorithm of CART and how the deviance statistics can be significant in the recursive binary splitting process of tree models, the algorithm can now be extended to accommodate data with Poisson response variables. Following the convention by (Lee and Jin, 2006), such trees have been given the name Poisson trees. Unlike the regression trees, the Poisson mean is the predicted value of the responses at the terminal nodes of the decision trees.

For a given Poisson regression, the Poisson deviance is given as

$$D_{\text{Poisson}} = 2 \sum_{y_i > 0} \left(y_i \log(\frac{y_i}{\hat{\lambda}_i}) - y_i - \hat{\lambda}_i \right) \, , \tag{3.6}$$

where $\hat{\lambda}_i$ is the predicted Poisson mean for y_i in a given node. Thus, by specifying the Poisson deviance in (3.5) for the respective parent and child nodes, the goodness

of split measure for the Poisson is defined.

3.5 ZIP trees

Zero-inflated Poisson (ZIP) tree is the conventional name that (Lee and Jin, 2006) gave to the decision tree model that was built for data with ZIP response variable. Other than using the deviance, (Lee and Jin, 2006) used a different statistic but similar to compute the goodness of split measure for the ZIP tree. This statistics was basically based on the likelihood ratio equality test of two sample means with the hypothesis, $H_0 : \mu_l = \mu_r$ verses $H_0 : \mu_l \neq \mu_r$, where μ_l and μ_r are the means in the left and right child nodes respectively. The likelihood ratio test statistic for the hypothesis was given as

$$\begin{aligned} \log(\lambda(y)) = \phi_{\text{ZIP}} =& [-\, max \log(L(y))] - [-\, max \log(L(y_l))] - \\ & [-\, max \log(L(y_r))] \geq 0 \,, \end{aligned} \tag{3.7}$$

where $max \log(L(y))$ is the maximum log-likelihood of the responses in the parent node. Equivalently, $max \ \log(L(y_l))$ and $max \ \log(L(y_r))$ correspond to the maximum log-likelihood in the left and right child nodes respectively. $\log(\lambda(y))$ was used as the goodness of split measure for the ZIP tree. From (2.7), the maximum log likelihood of the ZIP is given as

$$\begin{aligned} \log(L(y)) = & \sum_{y_i=0} \log\left[\hat{p} + (1-\hat{p})e^{-\hat{\lambda}}\right] + \sum_{y_i>0} \left[\log(1-\hat{p}) - \hat{\lambda}\right] + \\ & \sum_{y_i>0} y_i \log(\hat{\lambda}) - \sum_{y_i>0} \log(y_i!) \,, \end{aligned} \tag{3.8}$$

where $\hat{p}$ and $\hat{\lambda}$ are the maximum log-likelihood estimates of p and λ from the ZIP distribution of (2.6). Specifying $\log(L(y))$ in (3.7) for the respective parent and child

nodes, the goodness of split measure for the ZIP tree was defined. Thus, the ZIP maximum log-likelihood is computed for responses that fall within a particular node.

4 MZIP Decision Trees

4.1 Introduction

In this chapter, we introduce the main methodology of this thesis. We propose the MZIP decision tree model, having explained how the decision tree algorithm works in the previous chapters. The idea is to extend the ZIP tree model to rather predict the marginal mean with respect to the entire population instead of just the Poisson mean as is the case of the ZIP decision tree model.

4.2 Decision Tree for MZIP Data

In this thesis, the goal is to establish a decision tree method for MZIP regression models an alternative model in terms of decision tree. Following the naming convention, let's call it the MZIP decision tree. The algorithm to build the MZIP tree is based on the binary recursive splitting in CART and uses the maximum log-likelihood or the deviance of from the MZIP distribution as the goodness of split measure.

In this research, we use the homogeneous MZIP model to compute the goodness of split measure given in (3.5) for the MZIP tree. As in (2.10), we consider

$$\begin{aligned} \log(\frac{p}{1-p}) &= \gamma_0 \,, \\ log(v) &= \beta_0 \,, \end{aligned} \tag{4.1}$$

where β_0 and γ_0 are parameters. From (2.8) and (4.1), we have

$$e^{\beta_0} = \lambda \frac{e^{\gamma_0}}{1+e^{\gamma_0}} \quad \Leftrightarrow \quad \lambda = \frac{(1+e^{\gamma_0})e^{\beta_0}}{e^{\gamma_0}} \,. \tag{4.2}$$

Thus, $p = \frac{e^{\gamma_0}}{1+e^{\gamma_0}}$ and $1-p = \frac{1}{1+e^{\gamma_0}}$. Rewriting the likelihood function of the ZIP model in (2.7), we have

$$L(\beta_0, \gamma_0 | y_i) = \prod_{y_i=0} [\frac{1}{1+e^{\gamma_0}} + e^{-\lambda_i} \frac{e^{\gamma_0}}{1+e^{\gamma_0}}] \times \prod_{y_i>0} [\frac{e^{\gamma_0}}{1+e^{\gamma_0}} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}] \,. \tag{4.3}$$

By making appropriate substitution, the log-likelihood, $LL(\beta_0, \gamma_0 | y_i)$, from (4.3) becomes

$$\begin{aligned} LL(\beta_0, \gamma_0 | y_i) = &- \sum_{y_i} \ln(1+e^{\gamma_0}) + \sum_{y_i=0} \ln(1+e^{\gamma_0 - \lambda}) + \sum_{y_i>0} (\gamma_0 - \lambda) + \\ &\sum_{y_i>0} y_i(\beta - \gamma_0 + \ln(1+e^{\gamma_0})) - \sum_{y_i>0} \ln(y_i!) \,, \end{aligned} \tag{4.4}$$

where $\lambda = \frac{(1+e^{\gamma_0})e^{\beta_0}}{e^{\gamma_0}}$.

The values of β_0 and γ_0 that maximize the log-likelihood is found by solving the partial derivative equations

$$\frac{\delta LL}{\delta \beta_0} = \sum_{y_i=0} \frac{e^{\gamma_0 - \lambda}(-\frac{(1+e^{\gamma_0})e^{\beta_0}}{e^{\gamma_0}})}{1+e^{\gamma_0-\lambda}} + \sum_{y_i>0} [-\lambda + y_i] = 0 \tag{4.5}$$

and

$$\frac{\delta LL}{\delta \gamma_0} = -\sum_{\text{all i}} \frac{e^{\gamma_0}}{1+e^{\gamma_0}} + \sum_{y_i=0} \frac{e^{\gamma_0-\lambda}(1+e^{\beta_0-\gamma_0})}{1+e^{\gamma_0-\lambda}} + \sum_{y_i>0} [1+e^{\beta_0-\gamma_0} - y_i + \frac{y_i e^{\gamma_0}}{1+e^{\gamma_0}}] = 0\,. \tag{4.6}$$

From (4.5),

$$\begin{aligned} &\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}\lambda}{1+e^{\gamma_0-\lambda}}\right) + \frac{n_{y_i>0}}{N}\lambda = \bar{y} \\ &\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}}{1+e^{\gamma_0-\lambda}}\right) + \frac{n_{y_i>0}}{N} = \frac{\bar{y}}{\lambda}\,, \end{aligned}$$

where $\frac{n_{y_i=0}}{N}$ is the observed proportion of zero counts and $\frac{n_{y_i>0}}{N}$ is the observes proportion of non-zero counts.

From (4.6),

$$\begin{aligned} \Rightarrow -\frac{e^{\gamma_0}}{1+e^{\gamma_0}} + \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}}{1+e^{\gamma_0-\lambda}}\right)(1+e^{\gamma_0-\lambda}) + \frac{n_{y_i>0}}{N}(1+e^{\gamma_0-\lambda}) &= \frac{\bar{y}}{1+e^{\gamma_0}} \\ \Rightarrow -\frac{e^{\gamma_0}}{1+e^{\gamma_0}} + (1+e^{\gamma_0-\lambda})\left(\frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}}{1+e^{\gamma_0-\lambda}}\right) + \frac{n_{y_i>0}}{N}\right) &= \frac{\bar{y}}{1+e^{\gamma_0}} \\ \Rightarrow -\frac{e^{\gamma_0}}{1+e^{\gamma_0}} + (1+e^{\gamma_0-\lambda})\frac{\bar{y}}{\lambda} + \frac{n_{y_i>0}}{N} &= \frac{\bar{y}}{1+e^{\gamma_0}}\,, \end{aligned} \tag{4.7}$$

substituting $\lambda = \frac{(1+e^{\gamma_0})e^{\beta_0}}{e^{\gamma_0}}$ into the last line of (4.7) and solving implies that

$$e^{\beta_0} = \bar{y}\,. \tag{4.8}$$

Thus, $\bar{y}$ estimates the overall mean v and $\beta_0 = \ln(\bar{y})$ is the maximum likelihood estimate (m.l.e) of β_0 in (4.4). This implies that in (4.2)

$$\lambda = \frac{(1+e^{\gamma_0})\bar{y}}{e^{\gamma_0}} \Rightarrow \frac{\bar{y}}{\lambda} = \frac{e^{\gamma_0}}{1+e^{\gamma_0}} \Rightarrow \lambda = (e^{-\gamma_0}+1)\bar{y}\,. \tag{4.9}$$

Continuing for (4.5), we have

$$\frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}}{1+e^{\gamma_0-\lambda}}\right)+\frac{n_{y_i>0}}{N}=\frac{e^{\gamma_0}}{1+e^{\gamma_0}}$$
$$\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0}}{e^{\lambda}+e^{\gamma_0}}\right)+\frac{n_{y_i>0}}{N}=\frac{e^{\gamma_0}}{1+e^{\gamma_0}}$$

$$\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0-\lambda}}{1+e^{\gamma_0-\lambda}}\right)+\frac{n_{y_i>0}}{N}=\frac{e^{\gamma_0}}{1+e^{\gamma_0}}$$
$$\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{e^{\gamma_0}}{e^{\lambda}+e^{\gamma_0}}\right)+\frac{n_{y_i>0}}{N}=\frac{e^{\gamma_0}}{1+e^{\gamma_0}}$$
$$\Rightarrow \frac{n_{y_i=0}}{N}\left(\frac{1}{e^{(e^{-\gamma_0}+1)\bar{y}-\gamma_0}+1}\right)+\frac{n_{y_i>0}}{N}=\frac{1}{e^{-\gamma_0}+1}\,. \tag{4.10}$$

There is no closed form of explicit solutions of γ_0 in (4.10). Newton's method can be used to implicitly for γ_0 and gives the m.l.e of γ_0.

Using the estimated values of $\hat{\beta}_0$ and $\hat{\gamma}_0$ in (4.4), we can set up the goodness of split criteria using the deviance as discussed in (3.5) to build our MZIP decision tree model.

5 Simulations Studies

5.1 Introduction

In this chapter, we use simulated data to investigate the performance of the methodology of the MZIP tree models. In the simulation study, we investigate the type 1 error, power and accuracy of the MZIP trees developed in Chapter 4.

5.2 Simulation Results

Simulated data are based on the following MZIP model:

$$\begin{aligned} \log(\frac{p_i}{1-p_i}) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \, , \\ \log(v_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \, . \end{aligned} \tag{5.1}$$

In the first simulation scenario, to calculate the Type 1 error, thus, the proportion of trees that have no splits, we set $\beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 0$. A continuous noise variable $x_1 \sim N(0, \sigma^2)$, with three different values of the variance, $\sigma^2 = 0.01$, 1 and4 and a binary noise variable, $x_2 \sim$ Bernoulli(0.5) are considered. Meanwhile, we choose five γ_0 values: -2, 1, 0, 1 and 2. We generate 500 samples of fixed size 300 for each scenario. Since the true model is homogeneous there shouldn't be any spit ideally. Therefore, the proportion 500 trees that have at least one split is defined as

the Type 1 error of the MZIP tree model. We choose the fixed complexity parameter, cp = 0.01 when fitting the MZIP trees for all simulations. The probability values are given in the Table 5.1 below; From Table 5.1, the maximum probability of type 1

TABLE 5.1: *The Type 1 error rate under the MZIP tree models for simulated data with the two noise variables x_1 and x_2; Outliers with extreme predicted deviances are removed.*

	σ^2		
γ_0	0.01	1	4
-2	0.5042	0.4902	0.5328
-1	0.2268	0.2168	0.2287
0	0.0183	0.0182	0.0161
1	0.0040	0.0041	0.0101
2	0.0040	0.0121	0.0060

error was 0.5328 and the least was 0.004. Notice that the value of γ_0 determines the chance split in the MZIP tree models. Large γ_0 values, $(\gamma_0 \geq 0)$ correspond to large proportions of Poisson count, while smaller γ_0 values, $(\gamma_0 < 0)$ implies a large number of excess zeros. Hence when there are a big proportion of excess zeros, even if the model accommodates the homogeneous marginal mean, the tree still has about a half chance of making splits for the case of $\gamma_0 = -2$. Of course, when the proportion of excess zeros is 50% or less, $(\gamma_0 \geq 0)$, the Type 1 error is $< 2\%$ in the simulation regardless of the noise.

In the second simulation, we consider:

$$\begin{aligned} \log(\frac{p_i}{1-p_i}) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \,, \\ \log(v_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \,. \end{aligned} \tag{5.2}$$

where we have five choices for γ_0= -2, -1, 0, 1 and 2, four choices of β_2=-2, -1, 1 and 2, $\beta_0 = 1$, $\beta_1 = \gamma_1 = \gamma_2 = 0$. The noise variable $x_1 \sim N(0, \sigma^2)$ with fixed variance. The only variable $x_2 \sim$ Bernoulli(0.5) is used to calculate the power, that is, the

probability of the MZIP trees making at least I binary split and also the proportion of choosing x_2 as the primary splitting variable among these trees. The results are summarized in Table 5.2 and Table 5.3 respectively

TABLE 5.2: *Power in terms of proportion of trees which has at least 1 split under the MZIP tree methodology for 500 simulated datasets with x_1 as the noise variable and x_2 as the true variable; No outliers were removed.*

	β_2			
γ_0	-2	-1	1	2
-2	0.9880	0.8520	0.9980	1.0000
-1	0.9940	0.7100	0.9860	1.0000
0	0.9280	0.9560	0.9920	0.9920
1	0.9880	1.0000	0.9700	0.9100
2	1.0000	1.0000	0.9720	0.6900

TABLE 5.3: *Power in terms of proportion of MZIP trees which chose the true variable x_2 as primary split for 500 simulated datasets with x_1 as the noise variable; Outliers with extreme predicted deviances were removed*

	β_2			
γ_0	-2	-1	1	2
-2	0.9749	0.8500	0.8891	0.9486
-1	0.9627	0.7640	0.9937	0.9980
0	0.9680	0.9850	0.9980	1.0000
1	0.9938	0.9797	0.9375	1.0000
2	0.9879	0.9393	0.7278	1.0000

From Table 5.2, the least probability is 0.6900 which corresponds to a large effect of true variable to the marginal mean of a mixed population with a small portion of excess zeros and the highest is 1. This indicates that, most of time, the MZIP tree model was able to correctly identify non-homogeneous MZIP distribution. In Table 5.3, the least probability is 0.764 and the highest is 1; showing that the power of the MZIP model to identify the true variable for the primary splitting variable no smaller than 20% and is more than 80% for most scenarios.

Now, under the conditions of the second simulation, importance of the true variable x_2 and the noise variable x_1 is also summarized in Table 5.4; Importance of a variable is the proportion of reduction in deviance using that variable as the splitting variable. It can be seen from Table 5.4 that the true variable x_2 plays more significant role than the noise variable x_1. We also compared the average predicted deviance of the MZIP trees with the true deviance of the simulated dataset under the second simulation as shown in Table 5.5. It can be seen that most of the time, the average predicted deviance is quite close to the average of the true deviance except for $\gamma_0 = -2$ when there are about more than 80% excess zeros.

The simulation results so far have indicated the MZIP goodness of split criteria that was developed in chapter 4 performs well and would be able to properly identify a primary variable that is related to a zero-inflated Poisson response in a tree building process.

TABLE 5.4: *Comparison of importance of x_1 and x_2 from the MZIP trees for 500 simulated datasets with x_1 as the noise variable and x_2 as the noise variable. Outliers with extreme predicted deviances were removed*

	β_2 Values											
	-2	-2	-2	-1	-1	-1	1	1	1	2	2	2
γ_0	x1.imp	x2.imp	imp.sd	x1.imp	x2.imp	imp.sd	x1.imp	x2.imp	imp.sd	x1.imp	x2.imp	imp.sd
-2	0.0835	0.9165	0.0913	0.1601	0.8399	0.2282	0.2347	0.7653	0.1971	0.1389	0.8611	0.1326
-1	0.077	0.923	0.0824	0.1548	0.8452	0.2051	0.0919	0.9081	0.1029	0.0692	0.9308	0.0405
0	0.1113	0.8887	0.1488	0.0988	0.9012	0.1378	0.0653	0.9347	0.0518	0.0635	0.9365	0.0301
1	0.0763	0.9237	0.103	0.0845	0.9155	0.1348	0.1224	0.8776	0.2286	0.0637	0.9363	0.0299
2	0.0771	0.9229	0.1022	0.1121	0.8879	0.1873	0.3206	0.6794	0.4187	0.0632	0.9368	0.0289

TABLE 5.5: *Comparison of predicted deviance (the sum of deviance in the terminal nodes) and true deviance of the MZIP trees for 500 simulated datasets with x_1 as the noise variable and x_2 as the true variable; Outliers with extreme predicted deviances and number of splits are removed*

	β_2 Values															
	-2	-2	-2	-2	-1	-1	-1	-1	1	1	1	1	2	2	2	2
γ_0	tr.ave	pr.ave	tr.se	pr.se	tr.ave	pr.ave	tr.se	pr.se	tr.ave	pr.ave	tr.se	pr.se	tr.ave	pr.ave	tr.se	pr.se
-2	396.29	537.75	49.23	67.83	406.68	845.70	42.76	144.85	466.57	532.91	58.80	56.54	480.96	693.65	63.32	94.01
-1	608.99	671.76	44.73	41.75	684.25	915.81	46.84	52.81	806.71	826.51	56.63	50.82	846.26	871.02	62.26	56.00
0	780.84	782.66	41.36	39.84	946.05	950.03	39.28	35.21	1164.75	1161.55	44.16	43.36	1238.92	1236.12	49.03	48.06
1	842.60	839.58	31.70	31.74	1013.19	1009.92	25.58	25.85	1339.21	1335.76	28.24	28.71	1451.14	1448.11	35.14	35.55
2	834.48	837.01	30.34	30.55	998.37	1003.00	23.81	23.84	1358.35	1368.34	24.29	25.15	1490.67	1505.92	26.32	28.54

6 An Empirical Study

6.1 Introduction

In this chapter, we apply the methodology to a real-life data to build a decision tree. The data used is the German Socioeconomic Panel (GSOEP) data (1984–1995) (Riphahn, Wambach, and Million, 2003). The programming language used for the analysis is R with the RPART package.

6.2 Background of the Dataset

The German Socioeconomic Panel (GSOEP) data (1984–1995) is used for empirical analysis with the MZIP tree method. The data were collected based on annual face-to-face individual or computer-assisted personal interviews with household members aged 16 or over living in Germany for comprehensive information to measure stability and change in living conditions (Frick, 2006). The pooled subsample of the GSOEP data (1984–1994) includes 7293 German citizens, aged 26 through 65. After removing missing values, the subsample only includes years 1984–1988, 1991, and 1994 with 14,243 male observations and 13,083 female observations. The dependent variable is the number of doctor visits, `Dovics`, in the last 3 months right before the survey with 37.09% observations as zero and the mean across the whole sample is

3.18 with a standard deviation of 5.69. One key independent variable is the public insurance indicator, `Public`, which divides people into the group mandatorily insured by the public insurance (Public=1) and the group voluntarily (Public=0) with the proportions of 88.57% versus 14.33%. Among those with coverage of public insurance, about 2.12% purchased add-on insurance (Addon=1) which takes up 1.88% of the whole data; the rest did not purchase add-on insurance (Addon=0). The add-on insurance indicator is another key covariate. The age, degree of health satisfaction (using integer scales 0–10 meaning bad to well) and the household income are continuous covariates.

6.3 Application of the MZIP Tree to the GSOEP data

In this section, we apply the MZIP tree to the GSOEP data. A simple random sample of 500 from the dataset is used to fit the tree model. A pre-specified complexity parameter of 0.001 is used to prune the decision tree if necessary. A minimum number of 30 observations is set for each terminal node. The pruned tree structure for the data is shown in Fig. 6.1 below. The splitting criteria used to build the tree is based on the maximum log-likelihood of the MZIP distribution of the number of doctor visits of the individuals in the sample. It can be seen that the tree is very representative of the summary statistics described above. The overall mean number of doctor visits in the first node with 100% observations is approximately 4. From the tree, patients with voluntary public health insurance visit the hospital approximately just 2 times in the last three months before the survey. For patients with mandatory public health insurance, those with add-on insurance also visited the doctor approximately 2 times. Among patients with mandatory public health insurance who do not have add-on insurance, the males visited the doctor 5 times and the

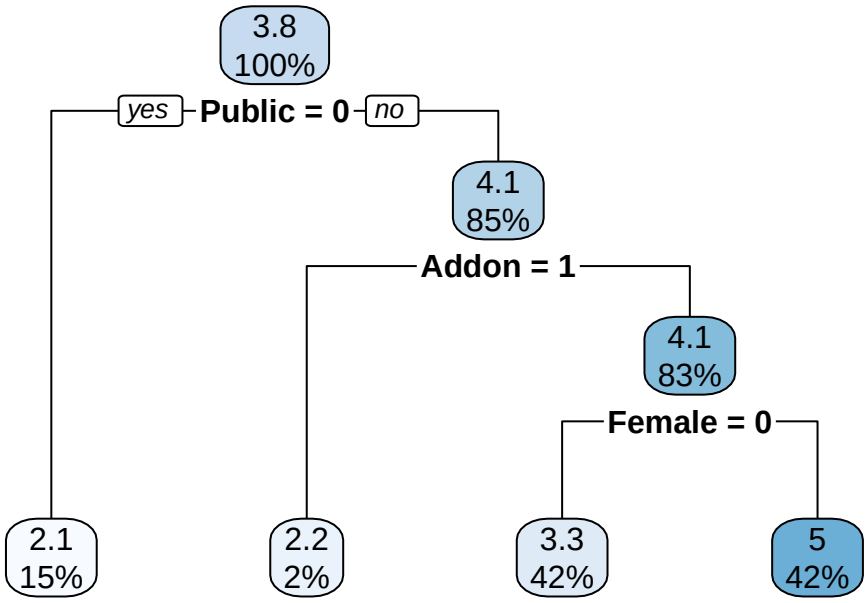

FIGURE 6.1: *Tree structure for the GSOEP data based on the MZIP goodness of split measure. The first number in each node is the predicted number of Doctor visits; percentage is the proportion of subjects in the given node*

females visited the doctor approximately 3 times. As indicated in the methodology, this tree model is very easy to interpret and very representative of the dataset. From the tree, the most important factor influencing the number of hospital visits is the public health insurance, followed by whether or not the patient had an add-on insurance, then gender. The variable importance plot of all the variables in considered in the tree building process is shown in Fig. 6.2. The variable importance measure basically depicts the amount of reduction in the weighted deviance at the various nodes due to splits over a given predictor. In Fig 6.2, a large proportion of the variable importance measure indicates an important predictor. The variable `Public` has the largest proportion of 43.07%, followed by `Addon`, with a proportion of 31.91%, then `Female`, with 18.54%; `Hhincome` with 2.64%; `Health` with 2.50% and the least, Age with a proportion of 1.34%.

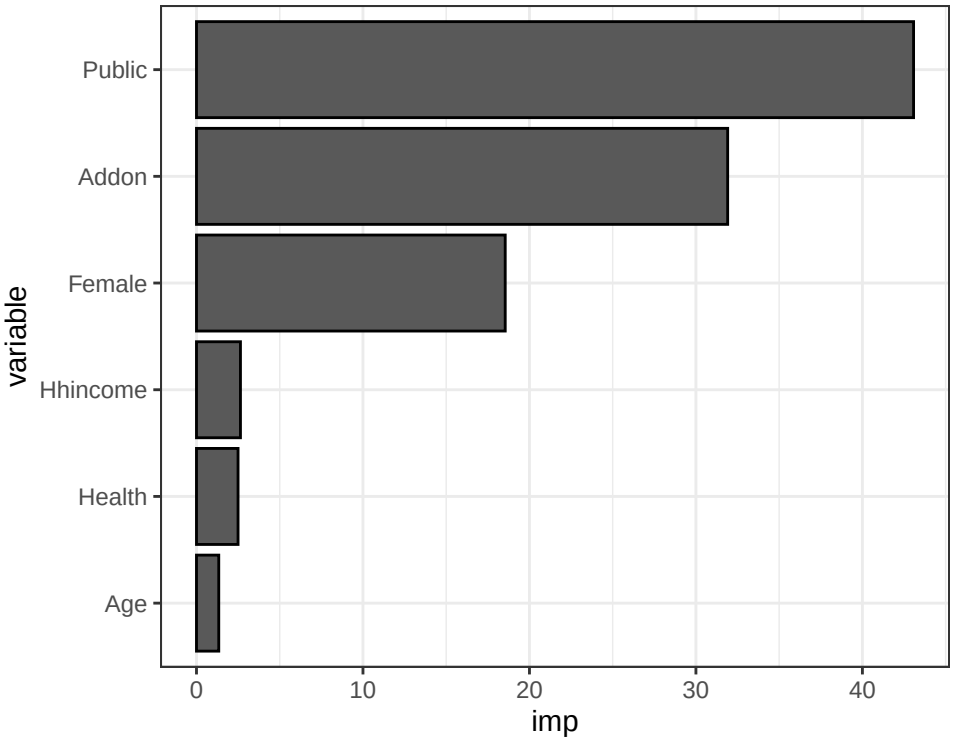

FIGURE 6.2: *Variable Importance plot of the MZIP tree in fig. 6.1; the horizontal axis (imp) represents the proportions of variable importance measure with a total of 100%*

7 Conclusion and Future Work

7.1 Introduction

In this final chapter, we summarize the research in the thesis and make some remarks about the MZIP tree model. We also discuss directions of future work.

7.2 The MZIP data and model

We have discussed count data and the various ways they are modeled. We have seen when the response count variable is zero-inflated, the options available include either the traditional zero-inflated model or the zero-inflated tree models. In this research, the focus was to the decision tree models. Tree models are known for their interpretability advantage and flexibility. In the literature, it is seen that there have been quite a number of works about the Poisson tree models and a little bit about the ZIP tree models with much reference to the work by (Lee and Jin, 2006). In this research we extended the idea of the ZIP tree model to develop the MZIP tree model.

The basic idea for the MZIP methodology for this research was to estimate the maximized MZIP log-ikelihood in (4.4) and use that to determine the goodness of split measure given by (3.5). In the tree building process, a binary recursive splitting

process is used where the split that results in the highest measure of the goodness of split is selected. The larger tree is then pruned by pre-specifying a complexity parameter which regularizes the complexity of the tree.

In chapter 5, we investigated the performance of the MZIP tree methodology using a variety of simulation studies. The results justified the performance of our methods. The MZIP model was then applied to a real-life healthcare dataset. The tree model was easy to interpret and was seen to be very representative of the summary statistics of the data.

7.3 Future work

In the process of developing the MZIP methodology in this research, it was realized that, the maximum log-likelihood of the ZIP and the MZIP could have the same value. However, in this research, it wasn't investigated how the MZIP tree and the ZIP tree model could perform in terms of misspecification where one is predicting the overall mean count while the other is predicting just the Poisson mean count.

In future research, careful study and comparison of the MZIP with other tree models for count data regarding performance in accuracy and robustness to misspecification can be considered.

Bibliography

Albert, J. M., W. Wang, and S. Nelson (2014). "Estimating overall exposure effects for zero-inflated regression models with application to dental caries." In: *Statistical Methods in Medical Research* 23, pp. 257–278.

Breiman, L. et al. (1984). "Classification and Regression Trees." In: *Computational Statistics and Data Analysis* 55.

Chaudhuri, P. et al. (1995). "Generalized regression trees." In: *Statistica Sinica* 5, pp. 641–666.

Crepon, B. and E. Duguet (1997). "Research and development, competition and innovation, pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity." In: *Journal of Econometrics* 79, pp. 355–378.

Frick, J. R. (2006). "A General Introduction to the German Socio-Economic Panel Study (SOEP)-Design, Contents and Data Structure (Waves A-V, 1984-2005)." In: *Deutsches Institut fur Wirtschaftsfor-schung, Berlin* 18, pp. 387–405.

Gupta, P. L., R. C. Gupta, and R. C. Tripathy (2004). "Score test for zero-inflated generalized Poisson model." In: *Communications in Statistics* 33, pp. 47–64.

Gurmu, S. (1997). "Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization." In: *Journal of Applied Econometrics* 12, pp. 225–242.

Heagerty, P. (1999). “Marginally specified logistic-normal models for longitudinal binary data.” In: *Biometrics* 55.3, pp. 688–698.

Heilbron, D. (1994). “Zero-altered and other regression models for count data with added zeros.” In: *Biometrical Journal* 12, pp. 531–547.

James, G. et al. (2013). “Introduction to Statistical learning with applications in R”. In: *Springer texts in Statistics* 103.

Kennan, J. (1985). “The duration of contract strikes in U.S. manufacturing”. In: *Journal of Econometrics* 28, pp. 5–28.

Lambert, D. (1992). “Zero-inflated Poisson regression, with an application to defects in manufacturing.” In: *Technometrics* 34, pp. 1–14.

Lee, K. et al. (2011). “Analysis of zero-inflated clustered count data: A marginalized model approach.” In: *Computational Statistics and Data Analysis* 55.1, pp. 824–837.

Lee, S. K. and S. Jin (2006). “Decision tree approaches for zero-inflated count data.” In: *Journal of Applied Statistics* 33, pp. 853–865.

Liu, X. et al. (2018). “Are marginalized two-part models superior to non-marginalized two-part models for count data with excess zeroes? Estimation of marginal effects, model misspecification, and model selection.” In: *Health Services and Outcomes Research Methodology* 18, pp. 175–214.

Loh, W. Y. and Y. S. Shih (1997). “Split selection methods for classification trees.” In: *Statistica Sinica* 7, pp. 815 –840.

Long, D. L. et al. (2014). “A marginalized zero-inflated Poisson regression model with overall exposure effects.” In: *Statistics in Medicine* 33.29, pp. 5151–5165.

Mathlouthi, W., M. Fredette, and D. Larocque (2015). “Regression trees and forests for non-homogeneous Poisson process.” In: *Statistics and Probability Letters* 96, pp. 204–211.

Miaou, S. P. (1994). "The relationship between truck accidents and geometric design of road sections. Poisson versus negative binomial regressions." In: *Accident Analysis and Prevention* 26, pp. 471–482.

Mullahy, J. (1986). "Specification and testing of some modified count data models." In: *Journal of Econometrics* 33, pp. 341–365.

Mullahy, J. (1997). "Heterogeneity, excess zeros, and the structure of count data models." In: *Journal of Applied Econometrics* 12, pp. 337–350.

Riphahn, R., A. Wambach, and A. Million (2003). "Incentive effects in the demand for health care: a bivariate panel count." In: *Journal of Applied Econometrics* 18, pp. 387–405.

Segal, M. R. (1992). "Tree-structured methods for longitudinal data." In: *Journal of the American Statistical Association* 87, pp. 407–418.

Therneau, T. M. and E. J. Atkinson (2004). "rpart: Recursive Partitioning;R package version 3.1-20". In: 55.

Therneau, T. M. and E. J. Atkinson (2019). "An Introduction to Recursive Partitioning using the RPART Routines". In: URL: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Welsh, A. et al. (1996). "Modeling the abundance of rare species– statistical-models for counts with extra zeros." In: *Ecological Modelling* 88, pp. 297–308.

Zimmermann, K. F. and J. Schwalbach (1991). "Determinanten der Patentak-tivitat". In: *Ifo-Studien* 37, pp. 201–227.

Vita

I, *Philip Amewudah*, registered as a graduate student at the Department of Mathematics, University of New Orleans in 2018 to pursue a masters degree in Mathematics. I was international student from Ghana and had my undergraduate degree from the Kwame Nkrumah University of Science and Technology, where I majored in BSc Statistics. At the Department of Mathematics at UNO, I concentrated mostly in Statistics related courses. I was also a graduate teaching assistant at the Department of Mathematics for the entire 2 years period of my master's program. I also worked as a research assistant with Dr Xueyan Liu in the 2019 summer semester. My research interest has been in Statistical Learning methods.