BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING
FAKULTA STROJNÍHO INŽENÝRSTVÍ

INSTITUTE OF MATHEMATICS
ÚSTAV MATEMATIKY

MODELS WITH TOUCHARD DISTRIBUTION

MODELY S TOUCHARDOVÝM ROZDĚLENÍM

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR
AUTOR PRÁCE
Michael Abimbola Ibukun

SUPERVISOR
VEDOUCÍ PRÁCE
doc. Mgr. Zuzana Hübnerová, Ph.D.

BRNO 2021

Assignment Master's Thesis

Institut:	Institute of Mathematics
Student:	**Michael Abimbola Ibukun**
Degree programm:	Applied Sciences in Engineering
Branch:	Mathematical Engineering
Supervisor:	**doc. Mgr. Zuzana Hübnerová, Ph.D.**
Academic year:	2020/21

As provided for by the Act No. 111/98 Coll. on higher education institutions and the BUT Study and Examination Regulations, the director of the Institute hereby assigns the following topic of Master's Thesis:

Models with Touchard Distribution

Brief Description:

Touchard distribution is a generalization of Poisson distribution allowing for under– or overdispersion. Moreover, under some conditions, this distribution still belongs to the family of distributions of exponential type.

Master's Thesis goals:

Study of the properties of Touchard distribution and selected models with this distribution.

Recommended bibliography:

MATSUSHITA, R., PIANTO, D., ANDRADE, B. B., CANCADO, A., SILVA, S. The Touchard distribution, Communications in Statistics - Theory and Methods. 2019, 48 (8), 2049-2059.

DOBSON, A. J., BARNETT, A. G. Introduction to Generalized Linear Models, 3rd. ed. 2008, Chapman and Hall.

Deadline for submission Master's Thesis is given by the Schedule of the Academic year 2020/21

In Brno,

L. S.

prof. RNDr. Josef Šlapal, CSc. Director of the Institute	doc. Ing. Jaroslav Katolický, Ph.D. FME dean

Abstrakt

Summary

In 2018, Raul Matsushita, Donald Pianto, Bernardo B. De Andrade, Andre Cançado & Sergio Da Silva published a paper titled ”Touchard distribution”, which presented a model that is a two-parameter extension of the Poisson distribution. This model has its normalizing constant related to the Touchard polynomials, hence the name of this model. This diploma thesis is concerned with the properties of the Touchard distribution for which δ is known. Two asymptotic tests based on two different statistics were carried out for comparison in a Touchard model with two independent samples, supported by simulations in R.

Klíčová slova

Keywords

Touchard distribution, Count data, Overdispersion, Underdispersion, Zero-Inflated counts, Exponential Family of distribution, Generalized Linear Models.

IBUKUN, M.*Modely s Touchardovým rozdělením.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2021. 53 s. Vedoucí doc. Mgr. Zuzana Hübnerová, Ph.D.

I declare that I have worked on this thesis independently under the supervision of doc. Mgr. Zuzana Hübnerová, Ph.D. and using the sources listed in the bibliography.

Michael Abimbola Ibukun

I would like to thank my supervisor doc. Mgr. Zuzana Hübnerová, Ph.D. for her indispensable guidance, support and helpful comments, without her patience and sacrifice of time I could not have completed this thesis. I thank all the lecturers who shared their knowledge with me in teaching. I thank all teaching and non-teaching staff of the Intermaths Erasmus Programme, my family and friends.

Michael Abimbola Ibukun

Contents

1. INTRODUCTION

The term ”count data” refers to a type of data in which the observations take only non-negative integer values; many data are in the form of counts. Examples include the census, the number of incidents, the number of arrivals and departures of flights at an airport, the number of patients arriving at a hospital due to a sickness and many more. The Poisson distribution is known to describe and model count data, which led to its high importance in the study of count data.

The Poisson model of count data is based on conditions which might be violated by some count data, such as equidispersion, which refers to the same mean and variance of the sample data. However, in many applications, the count data have overdispersion (a situation where the variance is greater than expected), underdispersion (a situation where the variance is less than expected), and excess zeros (a situation where the variation is greater than expected). For this reason, many generalizations of the Poisson distribution have been created, one of which is the Touchard distribution presented by Matsushita et al. 2018 in [2]. This model includes an additional parameter δ and a normalizing constant $\tau(\lambda, \delta)$ to help better fit the count data and limit the presence of excess zeros, overdispersion and underdispersion.

The Poisson distribution belongs to a family of distributions known as the exponential family. This family of distributions has special properties that are the basis of generalized linear models. In this thesis, I established that the Touchard distribution with δ is also known to belong to this family of distributions. The Touchard distribution is the same as the Poisson distribution if δ equals zero.

2. EXPONENTIAL FAMILY OF DISTRIBUTIONS

2.1. Introduction

In probability and statistics, the exponential family of distribution is a parametric set of probability distributions. This special form is chosen for mathematical convenience, based on some useful algebraic properties, as well as for generality. They are distinct because they posses a variety of desirable properties. Assume, we are given a single random variable Y whose probability distribution depends on a single parameter $\theta \in \Omega$, where Ω is an open set called the paramater space. The distribution belongs to the exponential family if its probability mass function (if Y discrete) or its density function (if Y continuous) can be written in the form

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

where a, b, s and t are known functions. The equation above can be rewritten in the form

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{2.1.1}$$

where $s(y) = \exp d(y)$ and $t(\theta) = \exp c(\theta)$

If $a(y) = y$, the distribution is said to be in the canonical form and $b(\theta)$ is sometimes called the natural parameter of the distribution. Our parameter of interest is θ, which is also called the canonical parameter (McCullagh and Nelder, 1989).

If there are other parameters, in addition to the parameter of interest θ, they are regarded as nuisance parameters forming parts of the functions a, b, c and d, and they are treated as though they are known.

Most of the commonly used distributions form the exponential family or subset of an exponential family, examples are: the Normal distribution, Exponential, Poisson, Gamma, Chi-squared, Beta, Geometric, Bernoulli distributions and many more.

Also, a number of common distributions are members of the exponential family, but only when certain parameters are fixed and known. For example: Binomial (with fixed number of trials), Negative binomial (with fixed number of failures) and also the Touchard distribution which will be discussed in this thesis. [1]

2.2. Mean and Variance of the Exponential Family

The mean and variance of the exponential family of distribution can also be given in a general form. This led to the expression of the expected value and variance of $a(Y)$.

Any probability density function $f(y;\theta)$ is normalized, i.e.

$$\int f(y;\theta)dy = 1 \tag{2.2.1}$$

in the continuous case,

$$\sum_{i=1}^{n} f(y_i;\theta) = 1$$

[1] The contents of this section are culled majorly from the book [1].

in the discrete case.

We apply the above result knowing that the order of integration and differentiation can be interchanged. This interchange is possible since the regularity conditions as discussed later in this text are satisfied.

If we differentiate both sides of (2.2.1) with respect to θ, we obtain

$$\frac{d}{d\theta}\int f(y;\theta)dy = \frac{d}{d\theta}\cdot 1 = 0$$

$$\int \frac{df(y;\theta)}{d\theta}dy = 0$$

$$\int \frac{d^2 f(y;\theta)}{d\theta^2}dy = 0$$

Applying this results for distributions in the exponential family. we have

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

$$\frac{df(y;\theta)}{d\theta} = [a(y)b'(\theta) + c'(\theta)]\, f(y;\theta)$$

$$\int [a(y)b'(\theta) + c'(\theta)]\, f(y;\theta)dy = 0$$

Therefore

$$b'(\theta)\mathrm{E}[a(y)] + c'(\theta) = 0$$

because

$$\int a(y)f(y;\theta)dy = \mathrm{E}[a(y)]$$

by the definition of the expected value and

$$\int c'(\theta)f(y;\theta)dy = c'(\theta)$$

So, we have

$$\mathrm{E}[a(Y)] = -c'(\theta)/b'(\theta)$$

Similarly, we obtain var$[a(Y)]$:

$$\frac{d^2 f(y;\theta)}{d\theta^2} = [a(y)b''(\theta) + c''(\theta)]\, f(y;\theta) + [a(y)b'(\theta) + c'(\theta)]^2\, f(y;\theta)$$

The second term on the right-hand side of the above equation can be rewritten as

$$\begin{aligned}[a(y)b'(\theta) + c'(\theta)]^2\, f(y;\theta) &= [b'(\theta)]^2 \left[a(y) + \frac{c'(\theta)}{b'(\theta)}\right]^2 f(y;\theta)\\ &= [b'(\theta)]^2\, \{a(y) - \mathrm{E}[a(Y)]\}^2 f(y;\theta)\end{aligned}$$

Therefore

$$\int \frac{d^2 f(y;\theta)}{d\theta^2}dy = b''(\theta)\mathrm{E}[a(Y)] + c''(\theta) + [b'(\theta)]^2\, \mathrm{var}[a(Y)] = 0$$

because $\int \{a(y) - \mathrm{E}[a(Y)]\}^2 f(y;\theta)dy = \mathrm{var}[a(Y)]$ by definition.

Finally,

$$\mathrm{var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

2.3. Maximum Likelihood Estimation

Let Ω denote the open set of all possible values of the parameter θ; Ω is called the parameter space. The maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.
Under the regularity conditions given later in this subsection, the following procedures in deriving the maximum likelihood estimate are asymptotically optimal. Assume, there is a random variable Y with the probability density function $f(y;\theta)$, where $\theta \in \Omega$. Consider, we have a random sample (iid) $Y_1 \dots Y_n$ from the distribution of Y. The parameter θ is unknown. The basis of our inferential procedures is the likelihood function given by

$$L(\theta;\mathbf{y}) = \prod_{i=1}^{n} f\left(y_i;\theta\right), \quad \theta \in \Omega \tag{2.3.1}$$

where $\mathbf{y} = (y_1, \dots, y_n)'$. Because we treat L as a function of θ we have transposed the y_i and θ in the argument of the likelihood function. It is easier and more convenient to use the logarithm of this function, so called log-likelihood, and we denote it by

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f\left(y_i;\theta\right), \quad \theta \in \Omega$$

Note that there is no loss of information in using $l(\theta)$ because the logarithm is a one-to-one function.

Regularity Conditions

It is well known, that the maximum likelihood estimators have favourable properties such as consistency or asymptotic efficiency as long as the so called regularity conditions are satisfied. This offer a theoretical justification for considering the mle.
Let θ_0 denote the true value of θ. It can be shown that the maximum of $L(\theta)$ asymptotically separates the true model at θ_0 from models at $\theta \neq \theta_0$. To prove this, the following regularity conditions must hold.

R0 The pdfs are distinct; i.e., $\theta \neq \theta' \Rightarrow f\left(y_i;\theta\right) \neq f\left(y_i;\theta'\right)$. The parameter identifies the pdf

R1 The pdfs have common support for all θ. i.e the support of Y_i does not depend on θ.

R2 The point θ_0 is an interior point in Ω.

R3 The pdf $f(y;\theta)$ is twice differentiable as a function of θ.

R4 The integral $\int f(y;\theta)dy$ can be differentiated twice under the integral sign as a function of θ.

R5 The pdf $f(y;\theta)$ is three times differentiable as a function of θ. Further, for all $\theta \in \Omega$, there exist a constant c and a function $M(y)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\ln f(y;\theta)\right| \leq M(y)$$

with $E_{\theta_0}[M(Y)] < \infty$, for all $\theta_0 - c < \theta < \theta_0 + c$ and all y in the support of Y

Note that conditions (R1) $-$ (R4) mean that the parameter θ does not appear in the endpoints of the interval in which $f(y;\theta) > 0$ and that we can interchange integration and differentiation with respect to θ.
Also, it can be shown, that under assumptions (R0) and (R1),

$$\lim_{n\to\infty} P_{\theta_0}\left[L\left(\theta_0, \mathbf{Y}\right) > L(\theta, \mathbf{Y})\right] = 1, \qquad \text{for all } \theta \neq \theta_0$$

A very detailed description of the regularity conditions can be found in [4].

Sufficient statistics and Factorization Theorem

As mentioned in [13], the concept of sufficiency arises as an attempt to answer the following question: Is there a statistic, i.e. a function $T\left(Y_1, \cdots, Y_n\right)$, that contains all the information in the sample about the parameter θ? If so, a reduction or compression of the original data to this statistic without loss of information is possible. The purpose of parameter estimation is to estimate the parameter θ from the random sample. It is also known that estimators can be expressed as a function of the random sample $Y_1, \cdots, Y_n$, such a function is called a statistic. Formally, any real-valued function $T = T\left(Y_1, \cdots, Y_n\right)$ of the observations in the sample is called a statistic.

If $T\left(Y_1, \cdots, Y_n\right)$ is a statistic and t is a particular value of T, then the conditional joint distribution of $Y_1, \cdots, Y_n$ given that $T = t$ can be calculated. In general, this joint conditional distribution will depend on the value of θ. Therefore, for each value of t, there will be a family of possible conditional distributions corresponding to the different possible values of θ in the parameter space Ω. However, it may happen that for each possible value of t, the conditional joint distribution of $Y_1, \cdots, Y_n$ given that $T = t$ is the same for all the values of $\theta \in \Omega$ and therefore does not actually depend on the value of θ. In this case, we say that T is a sufficient statistic for the parameter θ.

Definition 2.3.1. A statistic $T\left(Y_1, \cdots, Y_n\right)$ is said to be sufficient for θ if the conditional distribution of $Y_1, \cdots, Y_n$, given $T = t$, does not depend on θ for any value of t. In other words, given the value of T, we can gain no more knowledge about θ from knowing more about the probability distribution of $Y_1, \cdots, Y_n$. We could envision keeping only T and throwing away all the Y_i without losing any information.

It is difficult to determine if a given statistics T is sufficient or not, given definition (2.3.1), because of the difficulty in evaluating the conditional distribution. This led to the need for a simple method for finding a sufficient statistic which can be applied in many problems. This method is based on the following result, which was developed with increasing generality by R. A. Fisher in 1922, J. Neyman in 1935, and P.R. Halmos and L.J. Savage in 1949, and this result is known as the Factorization Theorem. [10]

Theorem 2.3.1 (Factorization Theorem)**.** Let $Y_1, \cdots, Y_n$ form a random sample from either a continuous distribution or a discrete distribution for which the probability density function or the probability mass function is $f(y \mid \theta)$, where the value of θ is unknown and belongs to a given parameter space Ω. A statistic $T(Y_1, \cdots, Y_n)$ is a sufficient statistic for θ if and only if the joint pdf or the joint probability mass function $f_n(\mathbf{y} \mid \theta)$ of $Y_1, \cdots, Y_n$ can be factorized as follows for all values of $\mathbf{y} = (y_1, \cdots, y_n) \in R^n$ and all values of $\theta \in \Omega$:

$$f_n(\mathbf{y} \mid \theta) = u(\mathbf{y})v[T(\mathbf{y}), \theta]$$

Here, the function u and v are nonnegative, the function u may depend on $\mathbf{y}$ but does not depend on θ, and the function v depends on θ but will depend on the observed value $\mathbf{y}$ only through the value of the statistic $T(\mathbf{y})$. In this expression, we can see that the statistic $T(Y_1, \cdots, Y_n)$ is like an "interface" between the random sample $Y_1, \cdots, Y_n$ and the function v.

Maximum Likelihood Estimate

Given the parameter space Ω, the maximum likelihood estimator of $\theta \in \Omega$ is the value $\widehat{\theta} = \widehat{\theta}(\mathbf{y})$ which maximizes the likelihood function, that is,

$$L(\widehat{\theta}; \mathbf{y}) \geq L(\theta; \mathbf{y})$$

for all θ in Ω. Also, $\widehat{\theta}$ is the value which maximizes the log-likelihood function $l(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y})$ since the logarithmic function is monotonic. Thus,

$$l(\widehat{\theta}; \mathbf{y}) \geq l(\theta; \mathbf{y})$$

for all θ in Ω.

To determine the MLE, we often find the extreme value of the log of the likelihood; that is, the MLE solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

2.4. Normal distribution

The normal distribution is a member of the exponential family, it has the probability density function defined as;

$$f(y; \mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

where μ is the parameter of interest and σ^2 is regarded as a nuisance parameter ($\sigma > 0$). This can be rewritten as

$$f(y; \mu) = \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right] \tag{2.4.1}$$

This is in the canonical form, since $a(y) = y$. The natural parameter is $b(\mu) = \mu/\sigma^2$ and the other terms in (1) are

$$c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right) \text{ and } d(y) = -\frac{y^2}{2\sigma^2}$$

(alternatively, the term $-\frac{1}{2}\log(2\pi\sigma^2)$ could be included in $d(y)$). The Normal distribution is used to model continuous data that have a symmetric distribution. It is widely used for three main reasons. First, many naturally occurring phenomena are well described by the Normal distribution; for example, height or blood pressure of people. Second, even if data are not Normally distributed (e.g., if their distribution is skewed) the average or total of a random sample of values will be approximately Normally distributed; this result is proved in the Central Limit Theorem. Third, there is a great deal of statistical theory developed for the Normal distribution, including sampling distributions derived from it and approximations to other distributions. For these reasons, if continuous data $\mathbf{y}$ are not Normally distributed it is often worthwhile trying to identify a transformation, such as $y' = \log y$ or $y' = \sqrt{y}$, which produces data y$'$ that are approximately Normal.

2.5. Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. [6]

The Poisson distribution is also a member of the exponential family of distribution, with the probability function for a discrete random variable Y given as

$$f(y, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where y takes the values $0, 1, 2, \ldots$ and $\lambda > 0$. This can be rewritten as

$$f(y, \lambda) = \exp(y \ln \lambda - \lambda - \ln y!) \tag{2.5.1}$$

$$\begin{aligned} a(y) &= y \\ b(\lambda) &= \ln \lambda \\ c(\lambda) &= -\lambda \\ d(y) &= -\ln y! \end{aligned}$$

which is in the canonical form because $a(y) = y$. Also the natural parameter is $\ln \lambda$. The Poisson distribution, denoted by $Y \sim \mathrm{Po}(\lambda)$, is used to model count data. Events such as the number of a product being purchased from a store each day, the number of jumps in a stock price in a given time interval, the number of vehicles passing through a toll gate between the early hours 5am and 8am, the number of laser photons hitting a detector in a particular time interval; may be modelled using the Poisson distribution.

Mean and Variance

$$\begin{aligned} \mathrm{E}[a(Y)] = \mathrm{E}[Y] &= -c'(\theta)/b'(\theta) \\ &= \lambda \\ \mathrm{Var}[a(Y)] = \mathrm{Var}[Y] &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \\ &= \lambda \end{aligned}$$

- One implication of the Poisson model is equi-dispersion. That is, the mean and variance are equal:
$$\mathrm{Var}\left[Y\right] = \mathrm{E}\left[Y\right]$$
- Overdispersion describes the observation that variation is higher than would be expected.
$$\mathrm{Var}\left[Y\right] > \mathrm{E}\left[Y\right]$$
- Underdispersion describes the observation that variation is lesser than would be expected.
$$\mathrm{Var}\left[Y\right] < \mathrm{E}\left[Y\right]$$

Maximum Likelihood Estimation

Let $Y = [Y_1, \ldots, Y_n]^T$ be independent random variables each with the Poisson distribution with the same parameter λ. The joint probability mass function(also the likelihood function) is given by

$$\begin{aligned} f\left(y_1, \ldots, y_n; \lambda\right) &= \prod_{i=1}^{n} f\left(y_i; \lambda\right) \\ &= \frac{\lambda^{\Sigma y_i} e^{-n\lambda}}{y_1! y_2! \ldots y_n!} \\ &= L\left(\lambda; y_1, \ldots, y_n\right) \end{aligned}$$

Finding the maximum likelihood estimate $\widehat{\boldsymbol{\lambda}}$

It is easier to find the maximum likelihood estimate using the log-likelihood function in the following steps.

$$\begin{aligned} l\left(\lambda; y_1, \ldots, y_n\right) &= \ln L\left(\lambda; y_1, \ldots, y_n\right) \\ &= \left(\sum_{i=1}^{n} y_i\right) \ln \lambda - n\lambda - \sum_{i=1}^{n} \left(\ln y_i!\right) \end{aligned}$$

By the derivative,

$$\frac{dl}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} y_i - n$$

At $\frac{dl}{d\lambda} = 0$, we obtain $\widehat{\lambda}$

$$\widehat{\lambda} = \sum_{i=1}^{n} Y_i / n = \bar{Y}$$

Since $d^2 l / d\lambda^2 = -\sum_{i=1}^{n} Y_i / \lambda^2 < 0, l$ has its maximum when $\lambda = \widehat{\lambda}$, confirming that $\bar{Y}$ is the maximum likelihood estimate of λ.

If a random variable has the Poisson distribution, its expected value and variance are equal. Real data that might be plausibly modelled by the Poisson distribution often have a larger variance and are said to be overdispersed, and the model may have to be adapted to reflect this feature.

2.6. Binomial distribution

Consider a series of binary events, called ”trials”, each with only two possible outcomes: ”success” or ”failure”. Let the random variable Y be the number of ”successes” in n independent trials in which the probability of success, π, is the same in all trials. Then Y has the Binomial distribution with probability density function

$$f(y;\pi) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{n-y}$$

where y takes the values $0, 1, 2, \dots, n$ and

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!}.$$

This is denoted by $Y \sim \mathrm{Bin}(n, \pi)$. Here π is the parameter of interest and n is assumed to be known. The probability function can be rewritten as

$$f(y;\pi) = \exp\left[y \ln \pi - y \ln(1-\pi) + n \ln(1-\pi) + \ln \left(\begin{array}{c} n \\ y \end{array} \right) \right]$$

which is of the form (2.1.1) with $b(\pi) = \ln \pi - \ln(1-\pi) = \ln[\pi/(1-\pi)]$. The Binomial distribution is usually the model of first choice for observations of a process with binary outcomes. Examples include the number of candidates who pass a test (the possible outcomes for each candidate being to pass or to fail) or the number of patients with some disease who are alive at a specified time since diagnosis (the possible outcomes being survival or death).

3. TOUCHARD DISTRIBUTION

The Touchard distribution is a generalization of the Poisson model to a two-parameter model which allows not only overdispersion or underdispersion, but excess zeros as well. This was inspired by the moments of the Poisson distribution, whose normalization constant relates to the Touchard polynomials. [7] [8][2]
Let Y be a random variable, whose probability mass function is defined as

$$f(y;\lambda,\delta)=\frac{\lambda^y(y+1)^\delta}{y!\tau(\lambda,\delta)} \tag{3.0.1}$$

where y takes the values $0, 1, 2, \ldots$, $\lambda > 0$ and $\delta \in \mathbb{R}$ are the distribution parameters, and the function

$$\tau(\lambda,\delta)=\sum_{j=0}^{\infty}\frac{\lambda^j(j+1)^\delta}{j!} \tag{3.0.2}$$

which normalizes the previous expression, is related to the Touchard polynomials (Rota [9] 1964, Chrysaphinou [8] 1985) and to the moment of order δ of a shifted Poisson distribution. This suggests $Y \sim$ Touchard (λ, δ), defined in (3.0.1), as a generalization of the Poisson distribution since for $\delta = 0, Y \sim$ Poisson (λ).

3.1. Properties

It can be seen that for known $\delta \in$ R, Touchard distribution is from the exponential family of distributions

$$\begin{aligned}
f(y;\lambda,\delta)&=\frac{\lambda^y(y+1)^\delta}{y!\tau(\lambda,\delta)}\\
&=\exp\left[\ln\left(\frac{\lambda^y(y+1)^\delta}{y!\tau(\lambda,\delta)}\right)\right]\\
&=\exp\left[\ln\lambda^y+\ln(y+1)^\delta-\ln y!-\ln\tau(\lambda,\delta)\right]
\end{aligned}$$

$$f(y;\lambda,\delta)=\exp[y\ln\lambda+\delta\ln(y+1)-\ln y!-\ln\tau(\lambda,\delta)] \tag{3.1.1}$$

Thus, the appropriate functions are:

$$\begin{aligned}
a(y)&=y\\
b(\lambda)&=\ln\lambda\\
c(\lambda)&=-\ln\tau(\lambda,\delta)\\
d(y)&=\delta\ln(y+1)-\ln y!
\end{aligned}$$

Using the properties of distributions from the exponential family, the mean and variance of the Touchard distribution can be expressed in terms of the normalizing function $\tau(\lambda, \delta)$ and as a multiple of λ.

$$\begin{aligned}
c'(\lambda)&=\frac{-1}{\tau(\lambda,\delta)}\sum_{j\in\mathbb{N}}j\frac{\lambda^{j-1}(j+1)^\delta}{j!}\\
b'(\lambda)&=\frac{1}{\lambda}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[a(y)] = \mathrm{E}[Y] &= -\frac{c'(\lambda)}{b'(\lambda)} \\
&= \frac{\lambda}{\tau(\lambda,\delta)} \sum_{j\in\mathbb{N}} \frac{j\lambda^{j-1}(j+1)^\delta}{j!} \\
&= \frac{1}{\tau(\lambda,\delta)} \sum_{j\in\mathbb{N}} \frac{j\lambda^{j}(j+1)^\delta}{j!} \qquad (3.1.2)\\
&= \frac{1}{\tau(\lambda,\delta)} \sum_{j\in\mathbb{N}} \left[\frac{\lambda^{j}(j+1)^{\delta+1}}{j!} - \frac{\lambda^{j}(j+1)^{\delta}}{j!}\right] \\
&= \frac{\tau(\lambda,\delta+1)-\tau(\lambda,\delta)}{\tau(\lambda,\delta)} \\
&= \frac{\tau(\lambda,\delta+1)}{\tau(\lambda,\delta)} - 1 \qquad (3.1.3)
\end{aligned}$$

The form (3.1.3) expresses the mean of the Touchard distribution in terms of λ and δ, without analytical representation of λ explicitly in terms of δ and the mean. This λ and mean dependency can be derived numerically and will be needed in the next sections. From (3.1.2) above, we can derive the mean of the Touchard distribution in another useful form.

$$\begin{aligned}
\mu = \mathrm{E}[Y] &= \frac{1}{\tau(\lambda,\delta)} \cdot \frac{\lambda}{\lambda} \cdot \sum_{Y=0}^{\infty} \left[(Y+1)\cdot \frac{\lambda^{Y+1}((Y+1)+1)^\delta}{(Y+1)!}\right] \\
&= \frac{\lambda}{\tau(\lambda,\delta)} \sum_{Y=0}^{\infty} \frac{\lambda^{Y}(Y+2)^\delta}{Y!} \\
&= \lambda \sum_{Y=0}^{\infty} \frac{(Y+2)^\delta}{(Y+1)^\delta} \frac{\lambda^{Y}(Y+1)^\delta}{Y!\tau(\lambda,\delta)} \\
&= \lambda \cdot \mathrm{E}\left[\left(\frac{Y+2}{Y+1}\right)^\delta\right] \qquad (3.1.4)
\end{aligned}$$

It can be seen that this result is equivalent to the mean of the Poisson distribution when $\delta = 0$, ans also that

$$\begin{cases} \mu > \lambda, & \delta > 0 \\ \mu < \lambda, & \delta < 0 \end{cases}$$

The variance of the Touchard distribution can be written from

$$E[Y] = -\lambda c'(\lambda) = \frac{\tau(\lambda,\delta+1)}{\tau(\lambda,\delta)} - 1 \qquad (3.1.5)$$

$$\begin{aligned}
\mathrm{Var}[Y] &= \frac{b''(\lambda)c'(\lambda) - c''(\lambda)b'(\lambda)}{[b'(\lambda)]^3} \qquad (3.1.6)\\
&= \frac{1}{b'(\lambda)} \frac{d}{d\lambda}(E[Y]) \qquad (3.1.7)
\end{aligned}$$

By differentiating both sides of (3.1.5) and substitution into (3.1.7), we have:

$$\begin{aligned}
\sigma^2 = \mathrm{Var}[Y] &= -\lambda\left[c'(\lambda) + \lambda c''(\lambda)\right] \\
&= \frac{\tau(\lambda,\delta)\tau'(\lambda,\delta+1) - \tau(\lambda,\delta+1)\tau'(\lambda,\delta)}{\tau^2(\lambda,\delta)} \\
&= \frac{\tau(\lambda,\delta)\tau(\lambda,\delta+2) - \tau(\lambda,\delta)\tau(\lambda,\delta+1) + \tau(\lambda,\delta)\tau(\lambda,\delta+1) - \tau^2(\lambda,\delta+1)}{\tau^2(\lambda,\delta)} \\
&= \frac{\tau(\lambda,\delta)\tau(\lambda,\delta+2) - \tau^2(\lambda,\delta+1)}{\tau^2(\lambda,\delta)} \\
&= \frac{\tau(\lambda,\delta+2)}{\tau(\lambda,\delta)} - \left[\frac{\tau(\lambda,\delta+1)}{\tau(\lambda,\delta)}\right]^2 \\
&= \lambda \mathrm{E}\left[(Y+1)\left(\frac{Y+2}{Y+1}\right)^{\delta}\right] - \mu^2
\end{aligned}$$

since

$$\tau'(\lambda,\delta) = \frac{\tau(\lambda,\delta+1) - \tau(\lambda,\delta)}{\lambda}$$

As can be seen in Figure (3.1), for higher values of μ(from the plot, $\mu > 20$), the dependency is almost linear, but for μ smaller than 20, there exists a non-linear dependence of the mean and variance, with some special behaviour for extremely low values of δ.

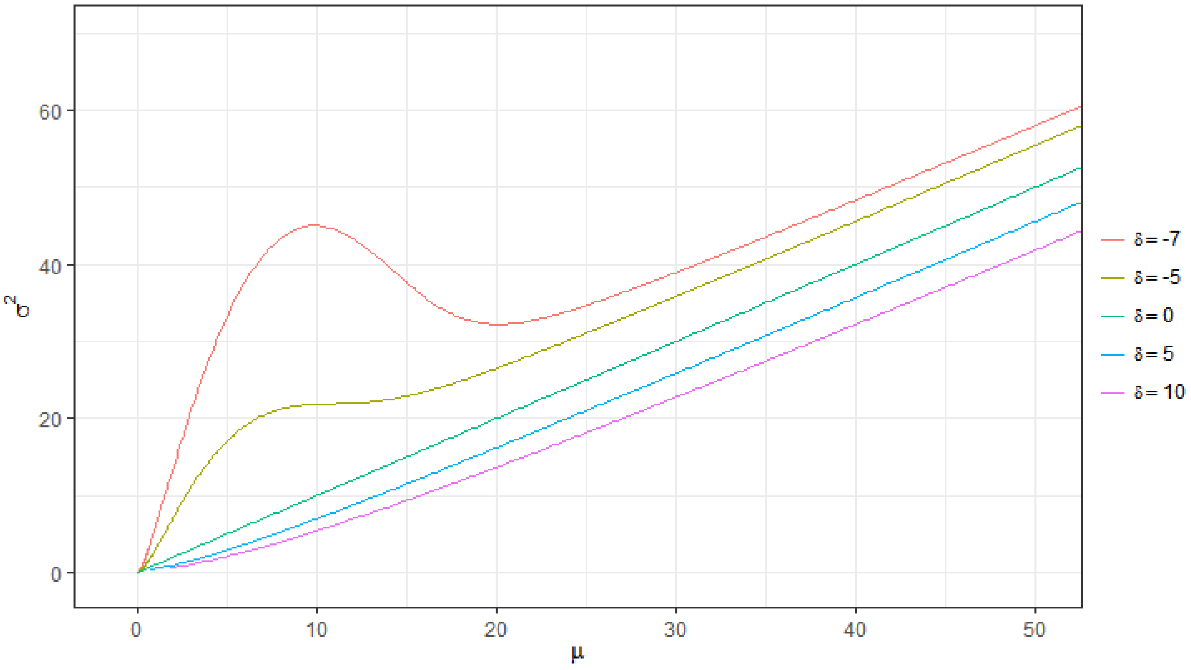


Figure 3.1: Variance-Mean dependency plot

3.2. Zero-Inflated Counts

Discrete probability distributions with a large probability mass at zero are said to be zero-inflated. These type of distributions are studies for example in [11, 14]. Excess zeros and zero-inflated counts are often encountered in the analysis of count data, particularly in relation to the Poisson distribution, but the term may be used in conjunction with any discrete distribution to indicate that there are more zeros than would be expected on the basis of the non-zero counts under assumed model. It is also possible for there to be fewer zero counts than expected, but this is much less common in practice. In statistics, a zero-inflated model is a statistical model based on a zero-inflated probability distribution, i.e. a distribution that allows for frequent zero-valued observations. One well-known zero-inflated model is Diane Lambert's zero-inflated Poisson model, which concerns a random event containing excess zero-count data in unit time.

Multiple models have been proposed in literature to model data with extra zeros as . The applications of these models are found in various disciplines from public health, economics, epidemiology, psychology, sociology, political sciences, agriculture, species abundance and road safety. It has been established that distributions such as the Poisson, Negative Binomial, Zero-inflated Poisson, Zero-inflated Negative Binomial, Generalized Poisson, Zero-inflated Generalized Poisson, Negative Binomial Lindley, Double Poisson and the Poisson Log-Normal have been used to fit count data with extra zeros.

For example, the number of insurance claims within a population for a certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim. The zero-inflated Poisson (ZIP) model mixes two zero generating processes. The first process generates zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero.

The Touchard probability mass function can be written by induction as:

$$\begin{aligned} f(y+1;\lambda,\delta) &= \frac{\lambda^{y+1}(y+2)^{\delta}}{(y+1)!\tau(\lambda,\delta)} = \frac{\lambda}{(y+1)} \cdot \frac{\lambda^{y}(y+2)^{\delta}}{y!\tau(\lambda,\delta)} \\ &= \frac{\lambda}{y+1} \frac{\lambda^{y}(y+1)^{\delta}}{y!\tau\left(\lambda,\delta\right)} \left[\frac{(y+2)^{\delta}}{(y+1)^{\delta}}\right] \\ &= \frac{\lambda}{y+1} \left(\frac{y+2}{y+1}\right)^{\delta} f(y;\lambda,\delta) \end{aligned}$$

Denote $f(y+1;\lambda,\delta)$ by T_{y+1} and $f(y;\lambda,\delta)$ by T_y, it can be seen that $T_{y+1}/T_y \downarrow 0$ as $y \uparrow +\infty$. Furthermore, the Touchard distribution naturally allows zero-inflated counts relative to the Poisson when λ and δ are chosen such that $T_{y^*} < T_0$ and $T_{y^*} < T_{y^*+1}$, for some fixed $y^* \geq 1$; i.e.[2]

$$\frac{T_{y^*}}{T_0} = \frac{\lambda^{y^*}\left(y^*+1\right)^{\delta}}{y^*!} < 1$$

and

$$\frac{T_{y^*+1}}{T_{y^*}} = \frac{\lambda}{y^*+1}\left(\frac{y^*+2}{y^*+1}\right)^{\delta} > 1$$

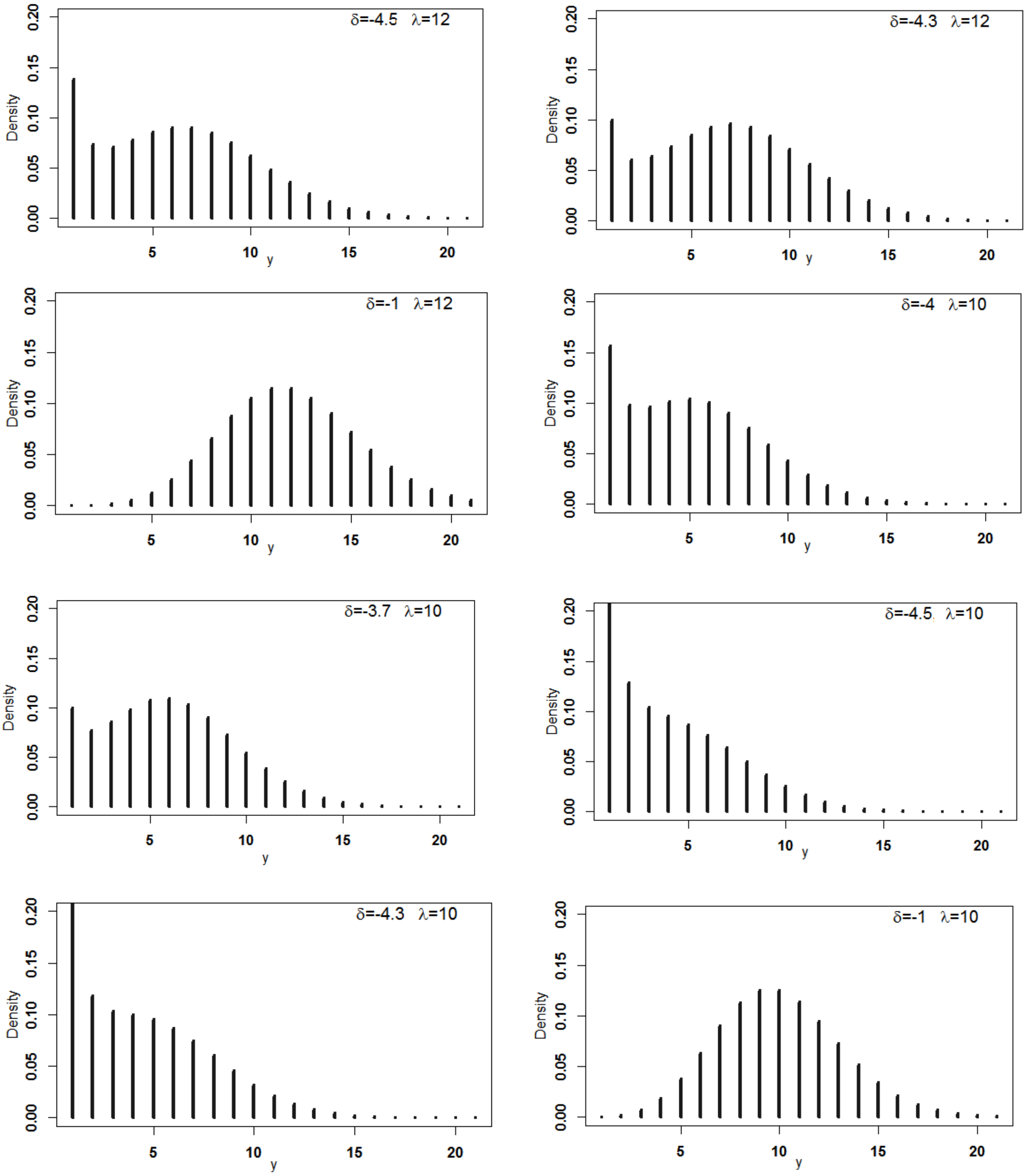


Figure 3.2: Examples of Touchard distributions, with $\lambda = 10, \lambda = 12$ and δ ranging from -4.0 to -1.0

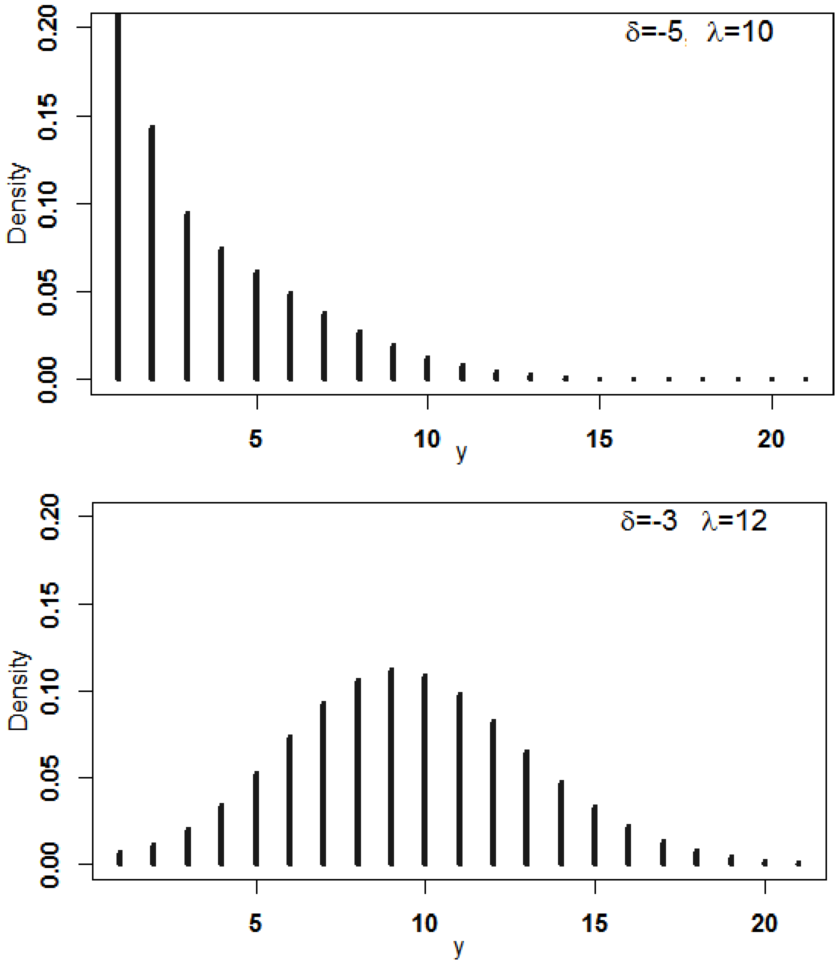


Figure 3.3: Examples of Touchard distributions, with $\lambda = 10, \lambda = 12$ and δ ranging from -4.0 to -1.0

Figures (3.2) and (3.3) show examples where in the Touchard distribution, excess zeros emerged when $\lambda = 10$ and $\delta = -4, -3.7$, also when $\lambda = 12$ and $\delta = -4.5, -4.3$

We can also consider a ratio of probabilities for $Y \sim \text{Touchard}(\lambda, \delta)$ and $X \sim \text{Poisson}(\lambda)$ being equal to 0. The ratio equals

$$P_0 = \frac{P(Y=0)}{P(X=0)} = \frac{e^{\lambda}}{\tau(\lambda, \delta)}$$

of the two probabilities at zero. This can be seen in figures (3.4) to (3.6), when $\delta = 0$ we have the Touchard distribution equal the Poisson distribution thereby $P_0 = 1$. With positive values of δ, the Touchard distribution has lesser probabilities of zeros compared to the Poisson distribution, and vice-versa when δ is a negative value.

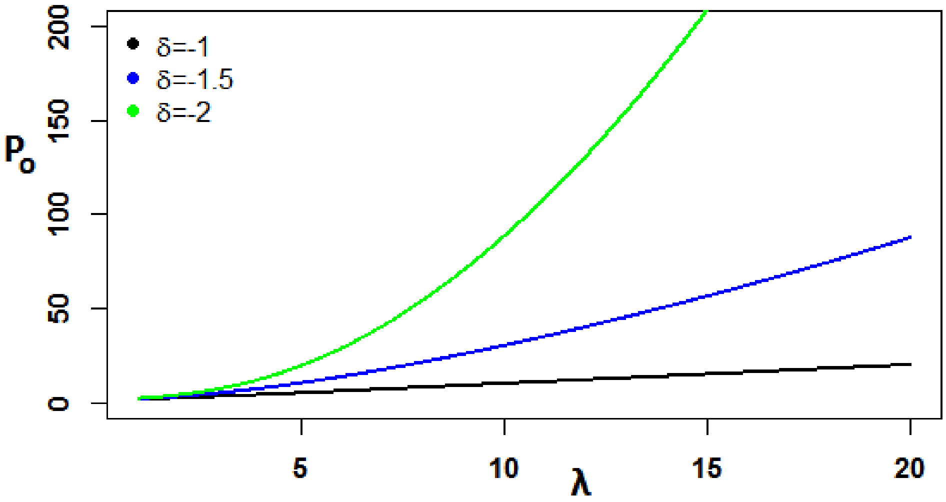


Figure 3.4: Ratio of Touchard and Poisson probabilities of zeros, at $\delta = -1, -1.5, -2$

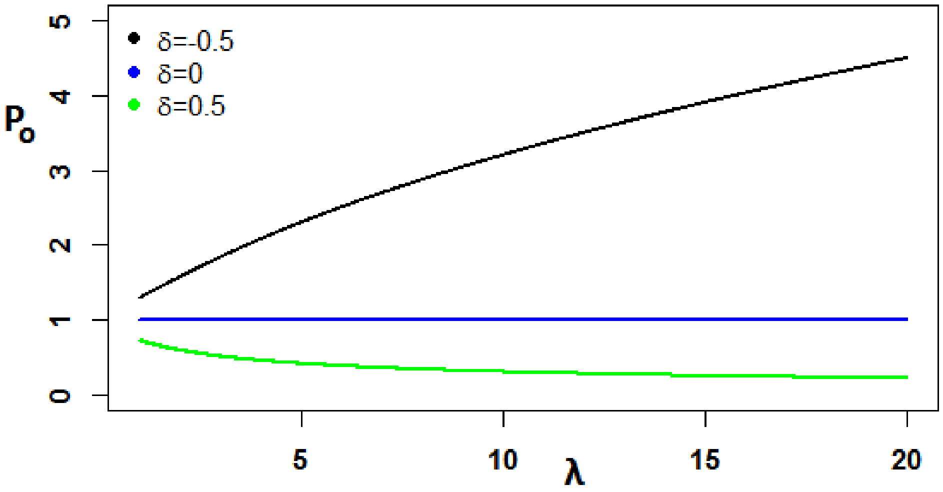


Figure 3.5: Ratio of Touchard and Poisson probabilities of zeros, at $\delta = -0.5, 0, 0.5$

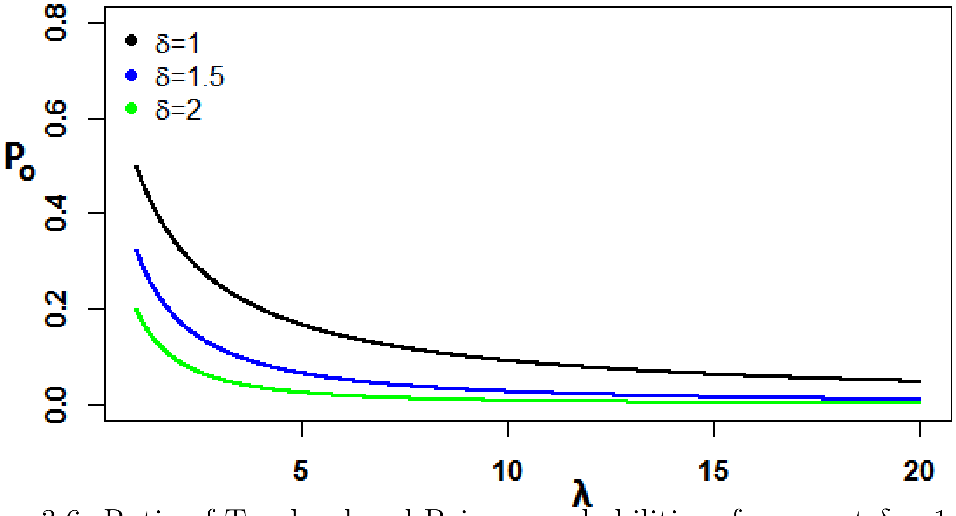


Figure 3.6: Ratio of Touchard and Poisson probabilities of zeros, at $\delta = 1, 1.5, 2$

3.3. Index of Dispersion

In probability theory and statistics, the index of dispersion also know by many other names such as the dispersion index, coefficient of dispersion, relative variance, or variance-to-mean ratio (VMR), the coefficient of variation, is a normalized measure of the dispersion of a probability distribution: it is a measure used to quantify whether a set of observed occurrences are clustered or dispersed compared to a standard statistical model. It is only defined when the mean μ is non-zero, and is generally only used for positive random variables, such as count data or time between events, or where the underlying distribution is assumed to be the exponential distribution, Weibull plot, Poisson distribution, etc. To assess the dispersion, we consider the ratio $r = \sigma^2/\mu$, which for Touchard distribution can be expressed as

$$r = \frac{E\left[(Y+1)\left(\frac{Y+2}{Y+1}\right)^{\delta}\right]}{\mathrm{E}\left[\left(\frac{Y+2}{Y+1}\right)^{\delta}\right]} - \mu$$

In the case where r is greater than 1, this describes overdispersion in the count data, and underdispersion when r is less than 1, but the case where $r = 0$ only exist for a constant with zero variance. Note that r above defines a dispersion index for counts.

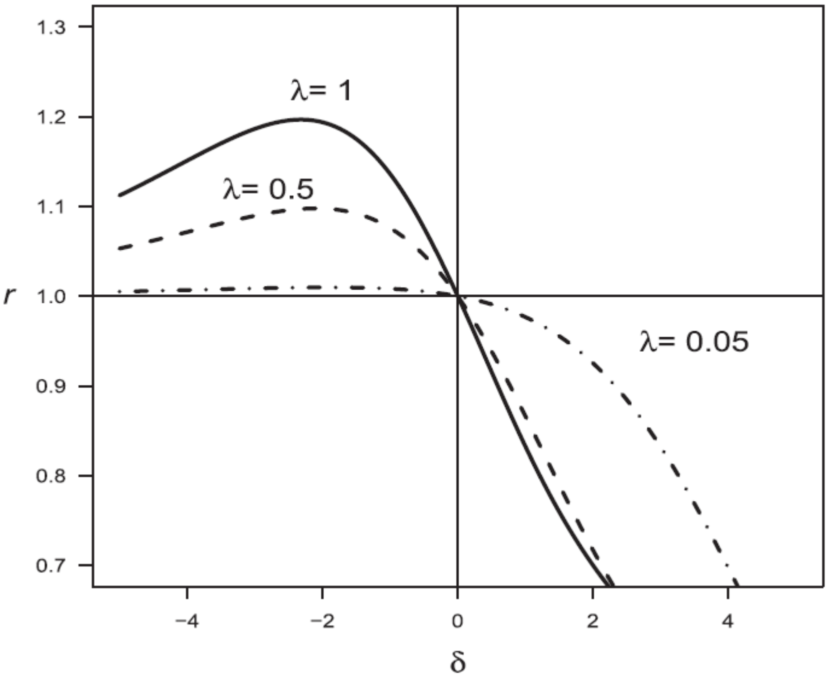


Figure 3.7: Behaviour of the ratio $r = \sigma^2/\mu$: overdispersion ($r > 1$) and underdispersion ($r < 1$).[2]

In the Poisson case ($\delta = 0$), we have $r = 1$. For $\delta > 0$, as $Y+1$ and $\{[Y+2]/[Y+1]\}^{\delta}$ are inversely (negatively) correlated, we have $r < 1$. Conversely, if $\delta < 0$, then $r > 1$.[2]

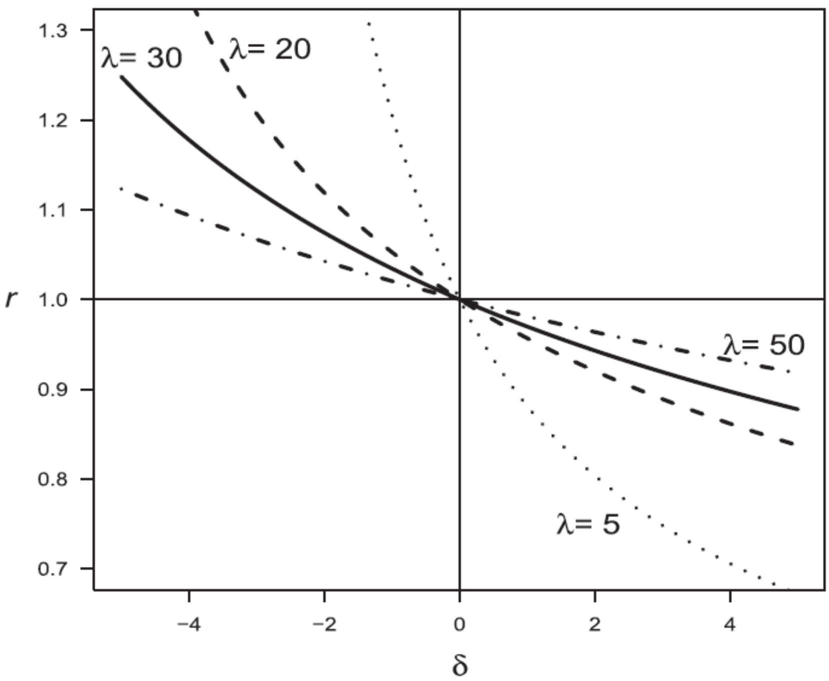


Figure 3.8: Behaviour of the ratio $r = \sigma^2/\mu$: overdispersion ($r > 1$) and underdispersion ($r < 1$).[2]

3.4. Skewness and Kurtosis

The r th moment of a Touchard random variable is described in [2] as a polynomial series of binomial type given by

$$\mathrm{E}\left[Y^r\right] = \sum_{j=0}^{r} \left(\begin{array}{c} r \\ j \end{array} \right) \frac{(-1)^{r-j}\tau(\lambda, \delta + j)}{\tau(\lambda, \delta)}$$

It was discussed in [5], the formulations of moments which include the uncorrected moments, moments about the mean, and the derivations of the moments about the mean from the uncorrected moments.

Definition 3.4.1 (Uncorrected Moments)**.** The expected value of Y^r for r any real number is termed the r th uncorrected (crude) moment (alternatively the r th moment about zero):

$$\mu_r'(Y) = \mu_r' = E\left[Y^r\right]$$

We will restrict r to only integer values.

Definition 3.4.2 (Moments about the Mean)**.** The r th moment about a constant a is $E\left[(Y-a)^r\right]$. When $a = \mu$, we have the r th moment about the mean (also called the r th central moment or the r th corrected moment),

$$\mu_r(Y) = \mu_r = E\left[(Y-\mu)^r\right] = E\left[(Y-E[Y])^r\right]$$

Moments about the Mean from Uncorrected Moments

It is often convenient to calculate the central moments μ_r from the uncorrected moments and, less often, vice versa. Formulas for this involve the binomial coefficients:

$$\mu_r = E\left[(Y - E[Y])^r\right] = \sum_{j=0}^{r} (-1)^j \begin{pmatrix} r \\ j \end{pmatrix} \mu'_{r-j}\mu^j$$

In particular

$$\begin{aligned} \mu_2 &= \mu'_2 - \mu^2 = \sigma^2 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu + 2\mu^3 \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4 \end{aligned}$$

Commonly used indices of the shape of a distribution are the moment ratios. The most important are

[1.] Index of skewness

$$\alpha_3(Y) = \sqrt{\beta_1(Y)} = \mu_3\left(\mu_2\right)^{-3/2}$$

[2.] Index of kurtosis

$$\alpha_4(Y) = \beta_2(Y) = \mu_4\left(\mu_2\right)^{-2}$$

The methods above are implored in the derivation of the moments of the Touchard distribution, this is used in solving for the indexes of skewness and kurtosis for different values λ and δ. This represented in figures (3.9) - (3.14). It can be seen that both reach the values of the normal distribution for high values of λ

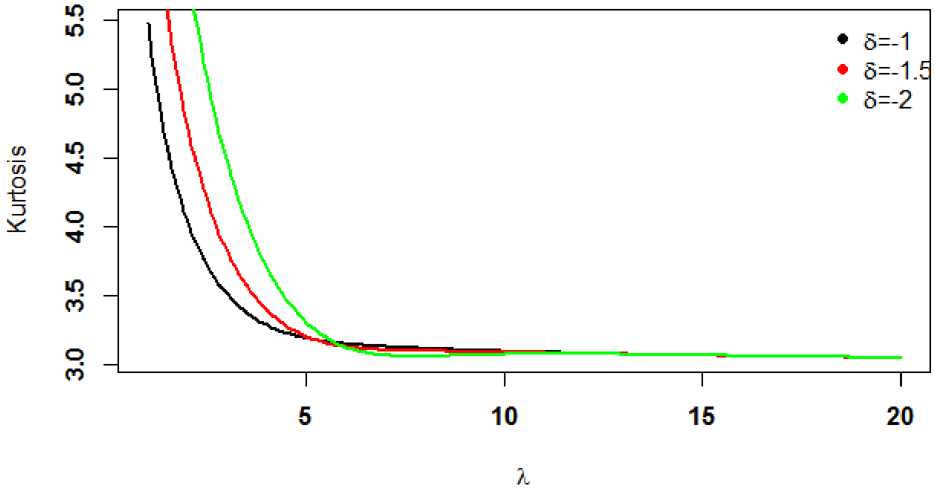


Figure 3.9: Plot of Kurtosis, at $\delta = -1, -1.5, -2$

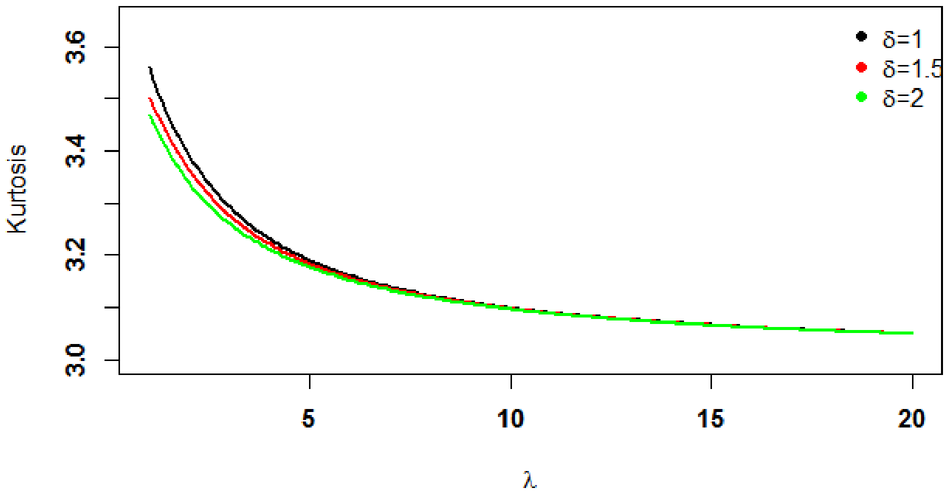


Figure 3.10: Plot of Kurtosis, at $\delta = 1, 1.5, 2$

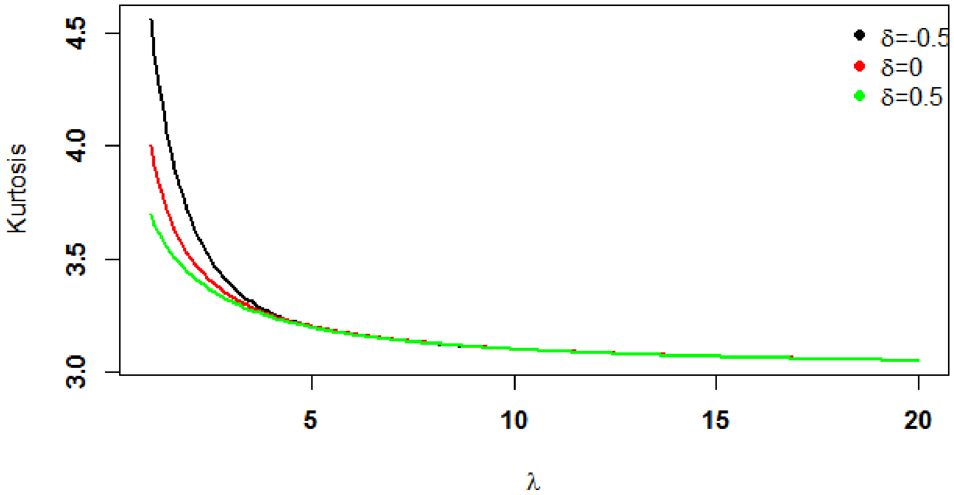


Figure 3.11: Plot of Kurtosis, at $\delta = -0.5, 0, 0.5$

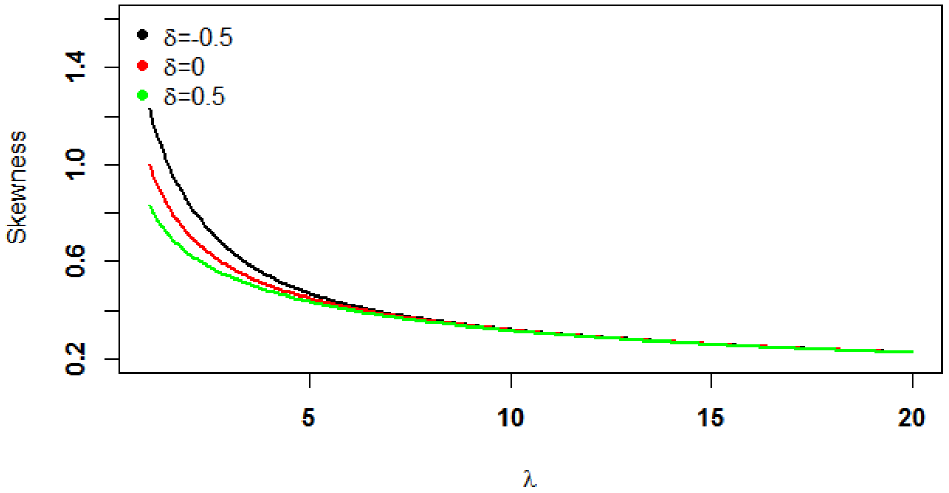


Figure 3.12: Plot of Skewness, at $\delta = -0.5, 0, 0.5$

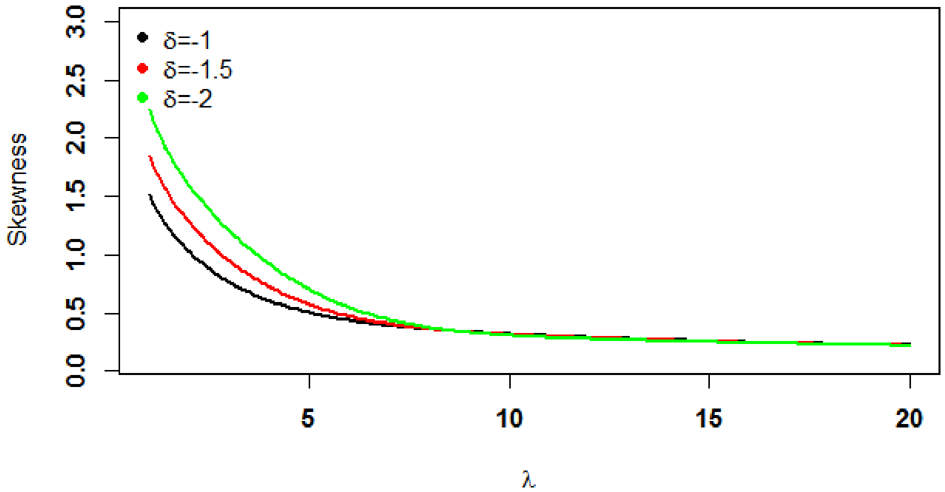


Figure 3.13: Plot of Skewness, at $\delta = -1, -1.5, -2$

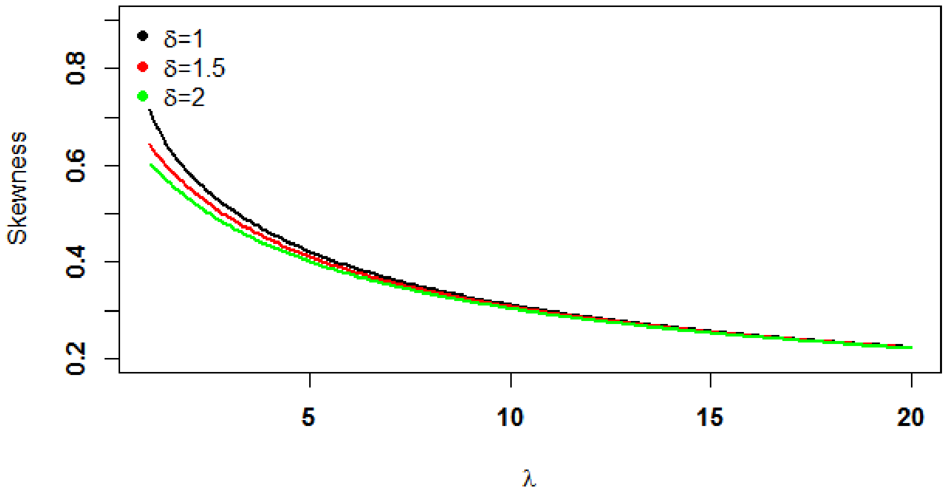


Figure 3.14: Plot of Skewness, at $\delta = 1, 1.5, 2$

3.5. Maximum Likelihood Estimation

Let $Y = [Y_1, \ldots, Y_n]^T$ be independent random variables each with the Touchard distribution with the same parameter λ and δ, let $y_1, \ldots, y_n$, be the n independent observations. Assume now, that both parameters λ and δ are unknown. The resulting likelihood can be written as

$$L(\lambda, \delta \mid y_1, \ldots, y_n) = \left(\prod_i y_i!\right)^{-1} \lambda^{S_1} \mathrm{e}^{\delta S_2} [\tau(\lambda, \delta)]^{-n} \tag{3.5.1}$$

with

$$S_1 = \sum_{i=1}^{n} Y_i$$

and

$$S_2 = \sum_{i=1}^{n} \ln(Y_i + 1)$$

being sufficient statistics by the factorization theorem.[2] To maximize the log-likelihood function $l(\lambda, \delta) = \ln L(\lambda, \delta \mid \{y_i\})$, the first and second derivatives of $\tau(\lambda, \delta)$ with respect

to λ and δ are needed.
First derivatives of $\tau(\lambda, \delta)$:

$$\frac{\partial\tau(\lambda,\delta)}{\partial\lambda} = \frac{\tau(\lambda,\delta)}{\lambda}\cdot\mu \tag{3.5.2}$$

$$\frac{\partial\tau(\lambda,\delta)}{\partial\delta} = \tau(\lambda,\delta)\mathrm{E}\{\ln[Y+1]\}. \tag{3.5.3}$$

Second derivatives of $\tau(\lambda, \delta)$:

$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\lambda^2} = \tau(\lambda,\delta)\cdot\frac{\mathrm{E}\left[Y^2\right]-\mu}{\lambda^2} \tag{3.5.4}$$

$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\delta^2} = \tau(\lambda,\delta)\mathrm{E}\left\{\ln^2[Y+1]\right\} \tag{3.5.5}$$

$$\frac{\partial^2\tau(\lambda,\delta)}{\partial\delta\partial\lambda} = \frac{\tau(\lambda,\delta)}{\lambda}\mathrm{E}\{Y\ln(Y+1)\} \tag{3.5.6}$$

By the use of equations (3.5.2) - (3.5.6), we obtain a system of two equations

$$\begin{cases} S_1 - n\mu = 0 \\ S_2 - n\mathrm{E}\{\ln[Y+1]\} = 0 \end{cases} \tag{3.5.7}$$

which are the maximum likelihood equations. Also, by the factorization theorem discussed in section (2.3), as the likelihood in (3.5.1) can be written as

$$L\left(\lambda,\delta \mid y_1,\ldots,y_n\right) = u(y_1,\ldots,y_n)v[S_1,S_2,\lambda,\theta],$$

maximizing $\ln L\left(\lambda,\delta \mid y_1,\ldots,y_n\right)$ with respect to λ and δ is equivalent to maximizing $v[S_1,S_2,\lambda,\theta]$ with respect to λ and δ. Therefore, the moments estimates of λ and δ, which satisfy (3.5.7), coincide with their corresponding maximum likelihood estimates.

In the case where δ is known, we maximize the log-likelihood function $l(\lambda,\delta) = \ln L\left(\lambda,\delta \mid y_1,\ldots,y_n\right)$ in one equation. In this case, only S_1 is the sufficient statistics. We have

$$l(\lambda) = \ln L = \ln\left[\exp(\delta S_2)\left(\prod_i y_i!\right)^{-1}\right] + \ln\left(\lambda^{S_1}\right) + \ln[\tau(\lambda,\delta)]^{-n}$$

and therefore

$$\frac{dl}{d\lambda} = \frac{S_1}{\lambda} - n\left[\frac{\tau(\lambda,\delta)}{\lambda}\cdot\mu\cdot\frac{1}{\tau(\lambda,\delta)}\right] = \frac{S_1-n\mu}{\lambda}$$

Also

$$\frac{d^2l}{d\lambda^2} = -\frac{S_1}{\lambda^2}$$

The log-likelihood function l is maximized at the stationary point for which $dl/d\lambda = 0$, where we have that $S_1 - n\mu = 0$. Given that $d^2l/d\lambda^2$ is always negative. The maximum likelihood estimate $\widehat{\mu}$ of μ is therefore $\bar{Y}$.

4. GENERALIZED LINEAR MODEL

4.1. Introduction

In this section, [1] already established theories that are used here. For a given set of random variables Y_i, $i = 1 \dots n$ which are independent, linear models are of the form

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}; \quad Y_i \sim \mathrm{N}\left(\mu_i, \sigma^2\right) \tag{4.1.1}$$

where the random variables Y_i for different subjects, indexed by the subscript i, may have different expected values μ_i.

The transposed vector $\mathbf{x}_i^T$ represents the ith row of the design matrix $\mathbf{X}$.

Advances in statistical theory and computer software allow us to use methods analogous to those developed for linear models in the following more general situations:

1. Response variables have distributions other than the Normal distribution, they may even be categorical rather than continuous.

2. Relationship between the response and explanatory variables need not be of the simple linear form in (4.1.1).

One of these advances has been the recognition that many of the "nice" properties of the Normal distribution are shared by a wider class of distributions called the exponential family of distributions discussed in section (2.1). A second advance is the extension of the numerical methods to estimate the parameters $\boldsymbol{\beta}$ from the linear model described in (4.1.1) to the situation where there is some non-linear injective function relating $\mathrm{E}(Y_i) = \mu_i$ to the linear component $\mathrm{x}_i^T \boldsymbol{\beta}$, that is

$$g(\mu_i) = \mathrm{x}_i^T \boldsymbol{\beta}$$

The function g is called the link function. In the initial formulation of generalized linear models by Nelder and Wedderburn (1972), g is a simple mathematical function.

4.2. Properties of GLMs

The generalized linear model is defined in terms of a set of independent random variables $Y_1, \dots, Y_N$, each with a distribution from the exponential family and the following properties:

1. The distribution of each Y_i has the canonical form and depends on a single parameter θ_i (the θ_i 's do not all have to be the same); thus,

$$f(y_i; \theta_i) = \exp\left[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\right].$$

2. The distributions of all the Y_i 's are of the same form (e.g., all Normal or all Binomial) so that the subscripts on b, c and d are not needed. Thus, the joint probability density function of $Y_1, \ldots, Y_N$ is

$$
\begin{aligned}
f\left(y_1, \ldots, y_N ; \theta_1, \ldots, \theta_N\right) & =\prod_{i=1}^N \exp \left[y_i b\left(\theta_i\right)+c\left(\theta_i\right)+d\left(y_i\right)\right] \\
& =\exp \left[\sum_{i=1}^N y_i b\left(\theta_i\right)+\sum_{i=1}^N c\left(\theta_i\right)+\sum_{i=1}^N d\left(y_i\right)\right]
\end{aligned}
$$

The parameters θ_i are typically not of direct interest (since there may be one for each observation). For model specification we are usually interested in a smaller set of parameters $\beta_1, \ldots, \beta_p$ (where $p<N$). Suppose that $\mathrm{E}\left(Y_i\right)=\mu_i$, where μ_i is some function of θ_i. For a generalized linear model there is a transformation of μ_i such that

$$
g\left(\mu_i\right)=\mathbf{x}_i^T \boldsymbol{\beta}
$$

where $\mathbf{x}_i^T \boldsymbol{\beta}$ is called the linear predictor and is often denoted by η_i.
In the above equation, we have the following properties:

i. g is a monotone, differentiable function called the link function; that is, it is flat, or increasing or decreasing with μ_i, but it cannot be increasing for some values of μ_i and decreasing for other values.

ii. The vector x_i is a $p \times 1$ vector of explanatory variables (covariates and dummy variables for levels of factors),

$$
\mathrm{x}_i=\left[\begin{array}{c}
x_{i1} \\
\vdots \\
x_{ip}
\end{array}\right] \text { so } \mathrm{x}_i^T=\left[\begin{array}{lll}
x_{i1} & \cdots & x_{ip}
\end{array}\right]
$$

iii. $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters $\boldsymbol{\beta}=\left[\begin{array}{c}\beta_1 \\ \vdots \\ \beta_p\end{array}\right]$.

The vector x_i^T is the i th row of the design matrix X. Thus, a generalized linear model has three components:

1. Response variables $Y_1, \ldots, Y_N$, which are assumed to share the same distribution from the exponential family;

2. A set of parameters β and explanatory variables

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & & x_{Np} \end{bmatrix}$$

3. A monotone link function g such that

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where

$$\mu_i = \mathrm{E}(Y_i)$$

4.3. Parameter estimates in GLM

Consider independent random variables $Y_1, \dots, Y_N$ in a generalized linear model. The parameters β in such a model are estimated by the maximum likelihood method. The log-likelihood function for all the Y_i 's is

$$l = \sum_{i=1}^{N} l_i = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i).$$

where the functions b, c and d are defined as it is in the exponential family.

To obtain the maximum likelihood estimator for the parameter β_j we need the derivatives of the log-likelihood with respect to the parameter. As stated in [1] the derivatives, called score statistics, are

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \left[\frac{\partial l_i}{\partial \beta_j} \right]$$

Finaly, by the chain rule we obtain

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right]$$

which is the score function U_j for Y_i. Then the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is the solution of the system of equations $U_j(\boldsymbol{\beta}) = 0$.

4.4. Poisson Regression

Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables, see [12]. As shown earlier, Poisson distribution is from the exponential family. So we can use GLM to model its mean as a function of vector of covariates $\mathbf{x}_i$. Thus we assume the response variables are independently Poisson distributed with

$$\mathrm{P}(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

Under the canonical link, in so called log-linear model, the mean number of events per period is given by

$$\mu_i = \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}$$

where $\boldsymbol{\beta}$ is a k-dimensional parameter. Observe that taking the exponential of the linear predictor ensures that the mean parameter μ_i is nonnegative.

Note that the variance of the Poisson random variable is equal to the mean

$$\mathrm{Var}(Y_i) = \mu_i$$

The equality of the mean and variance of Y_i is known as equidispersion. Also, the marginal effect of a regressor is given by

$$\frac{\partial \mu_i}{\partial x_{ij}} = \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}\beta_j = \mu_i \beta_j$$

Thus, a one-unit change in the j-th regressor leads to a proportional change in the conditional mean $\mathrm{E}(Y_i)$ of β_j. Poisson regression is estimated via maximum likelihood estimation. It usually requires a large sample size.

5. INFERENCE

Hypothesis tests in a statistical modelling framework are performed by comparing how well two related models fit the data. For generalized linear model, the two models should have the same probability distribution and the same link function, but the linear component of one model has more parameters than the other. The simpler model, corresponding to the null hypothesis H_0, must be a special case of the other more general model. If the simpler model fits the data as well as the more general model does, then it is preferred on the grounds of parsimony and H_0 is retained. If the more general model fits significantly better, then H_0 is rejected in favor of an alternative hypothesis H_1, which corresponds to the more general model. To make these comparisons, we use summary statistics to describe how well the models fit the data. These goodness of fit statistics may be based on the maximum value of the likelihood function, the maximum value of the log-likelihood function, the minimum value of the sum of squares criterion or a composite statistic based on the residuals. [1] The process and logic can be summarized as follows:

1. Specify a model M_0 corresponding to H_0. Specify a more general model M_1 (with M_0 as a special case of M_1).

2. Fit M_0 and calculate the goodness of fit statistic G_0. Fit M_1 and calculate the goodness of fit statistic G_1

3. Calculate the improvement in fit, usually $G_1 - G_0$ but G_1/G_0 is another possibility.

4. Use the sampling distribution of $G_1 - G_0$ (or some related statistic) to test the null hypothesis that $G_1 = G_0$ against the alternative hypothesis $G_1 \neq G_0$

5. If the hypothesis that $G_1 = G_0$ is not rejected, then H_0 is not rejected and M_0 is the preferred model. If the hypothesis $G_1 = G_0$ is rejected, then H_0 is rejected and M_1 is regarded as the better model.

For both forms of inference, sampling distributions are required. To calculate a confidence interval, the sampling distribution of the estimator is required. To test a hypothesis, the sampling distribution of the goodness of fit statistic is required.

For other distributions we need to rely on large-sample asymptotic results based on the Central Limit Theorem. The rigorous development of these results requires careful attention to various regularity conditions. For independent observations from distributions which belong to the exponential family and in particular for generalized linear models, the necessary conditions are indeed satisfied.

The basic idea is that under appropriate conditions, if S is a statistic of interest, then approximately

$$\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{var}(S)}} \sim \mathrm{N}(0, 1)$$

or equivalently

$$\frac{[S - \mathrm{E}(S)]^2}{\mathrm{var}(S)} \sim \chi^2(1)$$

where $\mathrm{E}(S)$ and $\mathrm{var}(S)$ are the expectation and variance of S, respectively. If there is a vector of statistics of interest $\mathbf{S} = \begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ with asymptotic expectation E(s) and asymptotic variance-covariance matrix $\mathbf{V}$, then approximately

$$[\mathbf{S} - \mathrm{E}(\mathbf{S})]^T \ \mathrm{V}^{-1}[\mathbf{S} - \mathrm{E}(\mathbf{S})] \sim \chi^2(p)$$

provided $\mathbf{V}$ is non-singular so a unique inverse matrix $\mathbf{V}^{-1}$ exists.

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated. This is called a saturated model. It is a generalized linear model with the same distribution and same link function as the model of interest.

5.1. Score statistics

Suppose $Y_1, \ldots, Y_N$ are independent random variables in a generalized linear model with parameters β, where $\mathrm{E}(Y_i) = \mu_i$ and $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$. The score statistics are

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \left[\frac{(Y_i - \mu_i)}{\mathrm{var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \qquad \text{for } j = 1, \ldots, p$$

As $\mathrm{E}(Y_i) = \mu_i$ for all i

$$\mathrm{E}(U_j) = 0 \qquad \text{for } j = 1, \ldots, p$$

The variance-covariance matrix of the score statistics is the information matrix $\mathfrak{I}$ with elements

$$\mathfrak{I}_{jk} = \mathrm{E}[U_j U_k].$$

The information matrix $\mathfrak{I}$ can also be written as

$$\mathfrak{I} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{X}$ is the design matrix and $\mathbf{W}$ is the $N \times N$ diagonal matrix with elements

$$w_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

If there is only one parameter β, the score statistic has the asymptotic sampling distribution

$$\frac{U}{\sqrt{\mathfrak{I}}} \sim \mathrm{N}(0,1), \text{ or equivalently } \frac{U^2}{\mathfrak{I}} \sim \chi^2(1)$$

because $\mathrm{E}(U) = 0$ and $\mathrm{var}(U) = \mathfrak{I}$. If there is a vector of parameters

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

then the score vector

$$\mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}$$

has the multivariate Normal distribution $\mathbf{U} \sim \mathrm{MVN}(\mathbf{0}, \mathfrak{I})$, at least asymptotically, and so

$$\mathbf{U}^T \mathfrak{I}^{-1} \mathbf{U} \sim \chi^2(p)$$

for large samples.

5.2. Log-likelihood ratio statistic

In general, let m denote the maximum number of parameters that can be estimated. Let $\beta_{\max}$ denote the parameter vector for the saturated model and $\mathrm{b}_{\max}$ denote the maximum likelihood estimator of $\boldsymbol{\beta}_{\max}$. The likelihood function for the saturated model evaluated at $\mathbf{b}_{\max}, L\left(\mathbf{b}_{\max}; \mathbf{y}\right)$, will be larger than any other likelihood function for these observations, with the same assumed distribution and link function, because it provides the most complete description of the data. Let $L(\mathbf{b}; \mathbf{y})$ denote the maximum value of the likelihood function for the model of interest. Then the likelihood ratio

$$\lambda = \frac{L\left(\mathbf{b}_{\max}; \mathbf{y}\right)}{L(\mathbf{b}; \mathbf{y})}$$

provides a way of assessing the goodness of fit for the model. In practice, the logarithm of the likelihood ratio, which is the difference between the log-likelihood functions,

$$\log \lambda = l\left(\mathbf{b}_{\max}; \mathbf{y}\right) - l(\mathbf{b}; \mathbf{y})$$

is used. Large values of $\ln \lambda$ suggest that the model of interest is a poor description of the data relative to the saturated model. To determine the critical region for $\ln \lambda$, we need its sampling distribution.

It has been established that $2 \ln \lambda$ has a chi-squared distribution. Therefore, $2 \ln \lambda$ rather than $\ln \lambda$ is the more commonly used statistic. It was called the deviance by Nelder and Wedderburn (1972).

The deviance, also called the log-likelihood (ratio) statistic, is therefore

$$D = 2\left[l\left(\mathbf{b}_{\max}; \mathbf{y}\right) - l(\mathbf{b}; \mathbf{y})\right]$$

where $\mathbf{b}$ is the maximum likelihood estimator of the parameter $\boldsymbol{\beta}$ (for which $\mathbf{U}(\mathbf{b}) = \mathbf{0}$). The sampling distribution of the deviance is approximately

$$D \sim \chi^2(m - p)$$

under the hypothesis of suitability of the considered model. Here, m is the number of parameters in the saturated model and p is the number of parameters in the model of interest. The deviance forms the basis for most hypothesis testing for generalized linear models.

Consider the null hypothesis

$$H_0 : \beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

corresponding to model M_0 and a more general hypothesis

$$H_1 : \beta = \beta_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

corresponding to M_1, with $q < p < N$. We can test H_0 against H_1 using the difference of the deviance statistics

$$\begin{aligned} \Delta D = D_0 - D_1 &= 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l\left(\mathbf{b}_0;\mathbf{y}\right)\right] - 2\left[l\left(\mathbf{b}_{\max};\mathbf{y}\right) - l\left(\mathbf{b}_1;\mathbf{y}\right)\right] \\ &= 2\left[l\left(\mathbf{b}_1;\mathbf{y}\right) - l\left(\mathbf{b}_0;\mathbf{y}\right)\right]. \end{aligned}$$

where b_0 is the estimate in the null hypothesis and b_0 the estimate in the alternative hypothesis. If both models describe the data well then $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$ so that $\triangle D \sim \chi^2(p-q)$, provided that certain independence conditions such as stated in the Cochran's theorem hold. If the value of $\triangle D$ is consistent with the $\chi^2(p-q)$ distribution we would generally choose the model M_0 corresponding to H_0 because it is simpler.

As stated in [4], the consequence of Cochran's theorem is that the difference of two independent random variables, $X_1^2 \sim \chi^2(m)$ and $X_2^2 \sim \chi^2(k)$, also has a chi-squared distribution

$$X^2 = X_1^2 - X_2^2 \sim \chi^2(m-k)$$

provided that $X^2 \geq 0$ and $m > k$.

6. COMPARISON OF TWO SAMPLES WITH TOUCHARD DISTRIBUTION

It has been shown in section (3.1) that the Touchard distribution(given that δ is known) is from the exponential family, this makes it possible to establish the properties of the generalized linear models and also make use of its methodologies for this distribution. Parameter estimation by maximum likelihood method and test of hypothesis can also be done on models with the Touchard distribution as it is for the generalized linear models, this makes it possible for us to evaluate the log-likelihood ratio test (also known as deviance) and the score statistics for the Touchard model in this chapter.

The method of Maximum Likelihood Estimate used in estimating the parameters of the Touchard distribution can also be modified to estimate the vector of unknown parameters, $\boldsymbol{\beta}$, of the linear predictor in the Touchard generalized linear model, and to show that the estimates are asymptotically normal. In this chapter, we focus on the test of hypothesis on the parameters of the Touchard model using the following statistics:

1. Score statistic

2. Log-likelihood ratio statistic (Deviance)

6.1. Model Description

Hypothesis tests in a statistical modelling framework are performed by comparing how well two related models fit the data. Here, we consider a simple model with two independent samples from the Touchard distribution

$$Y_{11}, \ldots, Y_{1N} \sim \text{Touchard}(\lambda_1, \delta)$$

$$Y_{21}, \ldots, Y_{2N} \sim \text{Touchard}(\lambda_2, \delta)$$

The hypothesis of interest is whether the parameters λ_1 and λ_2 are equal, i.e.

$$H_0 : \lambda_1 = \lambda_2 \Longleftrightarrow \mu_1 = \mu_2$$

We will test this hypothesis against

$$H_1 : \lambda_1 \neq \lambda_2 \Longleftrightarrow \mu_1 \neq \mu_2$$

The canonical link function

$$g\left(\mu_j\right) = x_j^T \beta = \eta_j$$

in the case of the Touchard model is defined in the following

$$\ln\left(\lambda_j\left(\mu_j\right)\right) = x_j^T \boldsymbol{\beta}$$

So in our model we can parametrize the linear predictor in the two samples as

$$\begin{aligned}\ln\left(\lambda_1\left(\mu_1\right)\right) &= \alpha+\beta_1\\ \ln\left(\lambda_2\left(\mu_2\right)\right) &= \alpha+\beta_2\end{aligned}$$

To obtain a full rank model we need to add a restriction that $\beta_1 = 0$. Then the equations are reduced to

$$\begin{aligned}\lambda_1 &= \exp(\alpha)\\ \lambda_2 &= \exp\left(\alpha+\beta_2\right) \qquad (6.1.1)\end{aligned}$$

In this case $\boldsymbol{\beta}^T = [\alpha\ \beta_2]$ and the vector of explanatory variables $x_i^T = [x_{i1}, x_{i2}]$.

And that gives the design matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} x_{111} & x_{112}\\ \vdots & \vdots\\ x_{1N1} & x_{1N2}\\ x_{211} & x_{212}\\ \vdots & \vdots\\ x_{2N1} & x_{2N2}\end{bmatrix} = \begin{bmatrix} 1 & 0\\ \vdots & \vdots\\ 1 & 0\\ 1 & 1\\ \vdots & \vdots\\ 1 & 1\end{bmatrix}$$

Using this parametrization we see that H_0

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \text{against} \quad \mathrm{H}_1 : \mu_1 \neq \mu_2$$

is equivalent to

$$\mathrm{H}_0' : \beta_2 = 0 \quad \text{against} \quad \mathrm{H}_1' : \beta_2 \neq 0$$

As it is in the generalized linear models, the test of hypothesis can be based on the Score statistics, Wald statistics or the log-likelihood ratio test. Here, we consider the Score statistics and the log-likelihood ratio test.

6.2. Maximum Likelihood Estimate

In section (4.3), we have seen that the maximum likelihood estimates are obtained by solving the equations $U(\beta_j) = 0$. The Touchard Model in this section is described by the equations: $\mathrm{E}\left(Y_{ji}\right) = \mu_j$ and $\lambda_j\left(\mu_j\right) = \exp(\mathrm{x}_j^T\beta) = \exp(\eta_j)$, where the variance of the Touchard distribution is expressed as $Var(Y_{ji}) = \sigma_j^2$. Therefore, the score function in the case of a model with two samples and k parameters is expressed as

$$U_k = \frac{\partial l}{\partial \beta_k} = \sum_{j=1}^{2}\sum_{i=1}^{N}\left[\frac{Y_{ji}-\mu_j}{\sigma_j^2}x_{jik}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)\right], \quad k = 1,2$$

Solving for $\partial\mu_j/\partial\eta_j$ from the equation $\lambda_j\left(\mu_j\right) = e^{\eta_j}$ we have

$$\frac{\partial \lambda_j\left(\mu_j\right)}{\partial \eta_j} = e^{\eta_j}$$

which is by the chain rule

$$\frac{\partial \lambda_j}{\partial \mu_j} \cdot \frac{\partial \mu_j}{\partial \eta_j} = e^{\eta_j}$$

therefore

$$\frac{\partial \mu_j}{\partial \eta_j} = e^{\eta_j} \frac{\partial \mu_j}{\partial \lambda_j}$$

By equation (3.1.7) in section (3.1) It has been established previously that $\sigma_j^2 = \lambda_j \frac{d\mu_j}{d\lambda_j}$, therefore, we have

$$\frac{d\mu_j}{d\lambda_j} = \frac{\sigma_j^2}{\lambda_j}$$

Thus, by substitution we arrive at the following

$$\frac{\partial \mu_j}{\partial \eta_j} = \lambda_j \left(\mu_j\right) \cdot \frac{\sigma_j^2}{\lambda_j \left(\mu_j\right)} = \sigma_j^2$$

This simplifies the score function into

$$U_k = \sum_{i=1}^{N} \sum_{j=1}^{2} \left[\frac{Y_{ji} - \mu_j}{\sigma_j^2} x_{jik} \sigma_j^2 \right]$$

$$U_k = \sum_{i=1}^{N} \sum_{j=1}^{2} \left[\left(Y_{ji} - \mu_j\right) x_{jik} \right], \quad k = 1, 2 \tag{6.2.1}$$

In solving for the difference of the deviance statistics ΔD, we need the estimates both under the null hypothesis and the alternative hypothesis.

The maximum likelihood estimate under the null hypothesis H_0

Let $\mu_1 = \mu_2 = \mu$, the maximum likelihood estimate under the null hypothesis H_0 can be derived from the score function as follows:

$$\frac{\partial l}{\partial \alpha} = U_1 = \sum_{i=1}^{N} \sum_{j=1}^{2} \left[\left(Y_{ji} - \mu_j\right) x_{ji1} \right] \quad = \sum_{i=1}^{N} \sum_{j=1}^{2} \left(Y_{ji} - \mu\right)$$

To find the maximum likelihood estimate of α (or μ) we set

$$\frac{\partial l}{\partial \alpha} = 0$$

which gives

$$\sum_{i=1}^{N} \sum_{j=1}^{2} Y_{ji} - 2N\hat{\mu} = 0$$

therefore

$$\widehat{\mu} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} Y_{ji}$$

The maximum likelihood estimate under the alternative hypothesis H_1 :

The maximum likelihood estimate in terms of the parameters α and β under the alternative hypothesis H_1 are solved for as follows:
For α : $\quad k = 1$

$$\begin{aligned}\frac{\partial l}{\partial \alpha} = U_1 &= \sum_{i=1}^{N}\sum_{j=1}^{2}\left[(Y_{ji}-\mu_j)\,x_{ji1}\right]\\ &= \sum_{i=1}^{N}\sum_{j=1}^{2}(Y_{ji}-\mu_j) \qquad \text{where } \mu_1 \neq \mu_2\end{aligned}$$

To find the maximum likelihood estimate of α (or μ_1) we set

$$\frac{\partial l}{\partial \alpha} = 0$$

which gives

$$\sum_{i=1}^{N}\sum_{j=1}^{2} Y_{ji} - N(\hat{\mu_1}+\hat{\mu_2}) = 0$$

$$\hat{\mu_1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} Y_{ji} - \hat{\mu_2} \tag{6.2.2}$$

For β_2 : $\quad k = 2$

$$U_2 = \frac{\partial l}{\partial \beta_2} = 0$$

implies

$$\sum_{i=1}^{N}\sum_{j=1}^{2}\left[(Y_{ji}-\mu_j)\,x_{ji2}\right] = 0,$$

$$x_{ji2} = \begin{cases} 1, \; j=2 \\ 0, \; j=1 \end{cases}$$

Therefore, $\sum_{i=1}^{N}\left[Y_{2i}-\hat{\mu_2}\right] = 0$ implies that

$$\hat{\mu_2} = \frac{1}{N}\sum_{i=1}^{N} Y_{2i} = \bar{Y_2} \tag{6.2.3}$$

By substituting (6.2.3) into (6.2.2),

$$\hat{\mu_1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} Y_{ji} - \bar{Y_2} = \bar{Y_1} \tag{6.2.4}$$

In section(3.1), it has been discussed in the properties of the Touchard distribution that λ cannot be expressed explicitly as a function of μ, therefore λ is solved for numerically using the R Software (https://cran.r-project.org/).

6.3. Score Statistics

The score statistics for the test of H_0' is described as

$$S = \mathbf{U}(\widehat{\lambda})^T \mathfrak{I}^{-1}(\widehat{\lambda})\mathbf{U}(\widehat{\lambda})$$

where $\widehat{\lambda}$ is the estimate under the null hypothesis and the diagonal matrix $\mathbf{W}$ has the elements

$$w_{ii} = \frac{1}{\operatorname{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Under the canonical link

$$\frac{\partial \mu_i}{\partial \eta_i} = \sigma_i^2$$

therefore

$$w_{ii} = \frac{(\sigma_i^2)^2}{\sigma_i^2} = \sigma_i^2$$

The resulting information matrix

$$\mathfrak{I} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

Using the design matrix of our model given earlier

$$\mathfrak{I} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 & \cdots & & 0 & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & 0 & \sigma_1^2 & 0 & & \vdots \\ 0 & & 0 & \sigma_2^2 & 0 & \vdots \\ \vdots & & & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

Then we obtain

$$\mathfrak{I} = \begin{bmatrix} N(\sigma_1^2 + \sigma_2^2) & N\sigma_2^2 \\ N\sigma_2^2 & N\sigma_2^2 \end{bmatrix}$$

Its value under the null hypothesis, when $\widehat{\sigma}^2 = \widehat{\sigma}_1^2 = \widehat{\sigma}_2^2$ is

$$\mathfrak{I}(\widehat{\sigma}^2) = \widehat{\sigma}^2 \begin{bmatrix} 2N & N \\ N & N \end{bmatrix}$$

To calculate the score statistics, we need the inverse

$$\mathfrak{I}(\widehat{\sigma}^2)^{-1} = \frac{1}{\widehat{\sigma}^2 N} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

Since $U(\widehat{\lambda}) = [0 \; U_2(\widehat{\lambda})]$, S can be written as

$$S = \frac{1}{\widehat{\sigma}^2 N} [0 \; U_2(\widehat{\lambda})] \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} [0 \; U_2(\widehat{\lambda})]^T$$

The score statistics is chi-square distributed with degree of freedom $p - q = 1$, which has an asymptotic distribution under the null hypothesis H_0.

6.4. Deviance

We test the null hypothesis H_0 against the alternative hypothesis H_1 using the difference of the deviance statistics which is to compare the maximized likelihoods of the two models

$$\Delta D = 2\left[l\left(\mathbf{b}_1;\mathbf{y}\right) - l\left(\mathbf{b}_0;\mathbf{y}\right)\right].$$

where $\mathbf{b}_0$ is the estimate in the null hypothesis and $\mathbf{b}_1$ the estimate in the alternative hypothesis. Note that in our model

$$l(\lambda_1,\lambda_2,\mathbf{y}) = \sum_{j=1}^{2} \ln\left[\exp(\delta S_{2j})\left(\prod_i y_{ji}!\right)^{-1}\right] + \sum_{j=1}^{2} \ln\left(\lambda^{S_{1j}}\right) + \sum_{j=1}^{2} \ln[\tau(\lambda_j,\delta)]^{-N}$$

If both models describe the data well, $\triangle D \sim \chi^2(1)$. If the value of $\triangle D$ is consistent with the $\chi^2(1)$ distribution we would generally choose the model M_0 corresponding to H_0 because it is simpler.

6.5. Results from simulations in R Software

In this section, I performed simulations based on our previously described model in the R Software [version 3.6.0 (2019-04-26), https://cran.r-project.org/]. We have our model with two sampling distributions:

$$Y_{11}, \dots, Y_{1N} \sim \text{Touchard}(\lambda_1, \delta)$$

$$Y_{21}, \dots, Y_{2N} \sim \text{Touchard}(\lambda_2, \delta)$$

where each sample is of size N, the link function described by

$$\begin{aligned} \lambda_1 &= \exp(\alpha) \\ \lambda_2 &= \exp\left(\alpha + \beta_2\right) \end{aligned} \tag{6.5.1}$$

The test for the hypothesis

$$\text{H}_0 : \mu_1 = \mu_2 \quad \text{against} \quad \text{H}_1 : \mu_1 \neq \mu_2$$

which is equivalent to

$$\text{H}_0' : \beta_2 = 0 \quad \text{against} \quad \text{H}_1' : \beta_2 \neq 0$$

is performed by the log-likelihood ratio test and the score statistic in 1000 simulations for each value of $\delta, \lambda_1, \lambda_2$. Parameters λ_1 and λ_2 are linked by a positive value h through

$$\lambda_2 = h\lambda_1$$

From (6.5.1) we have

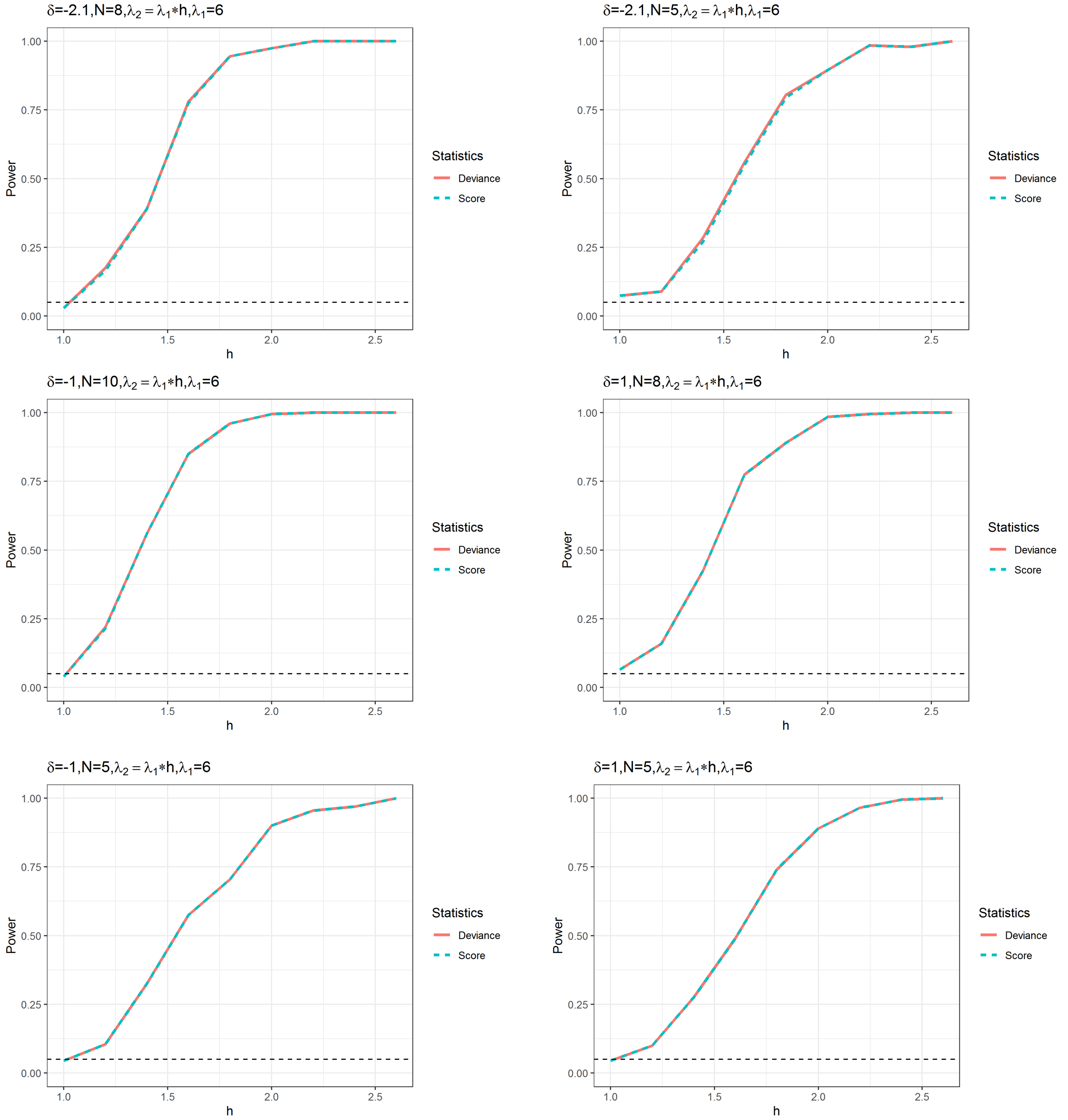
$$\lambda_2 = \lambda_1 \exp(\beta_2)$$

which implies $h = \exp(\beta_2)$.

The maximum likelihood estimates of μ in the case of the null hypothesis and μ_1, μ_2 in the case of the alternative hypothesis are calculated by using the score function. Since λ cannot be explicitly solved with respect to μ in the formulations and properties of the Touchard distribution, I implemented a numerical solution for each case in which λ is to be evaluated. I varied the value of δ in both positive and negative intervals in R close to zero, where errors occurred for some values of δ, such as $\delta = -3, -3.5, -4.5, -5.5, -7, \dots$ regardless of the values of N, h and λ. This is a consequence of the high probability of zeros in the samples as δ increases negatively. I programmed the computation of the score statistics and the log-likelihood ratio test based on the values of the maximum likelihood estimates in the hypothesis and compared them with the 95% quantile of χ^2 distribution with one degree of freedom (since the difference in the number of parameters in the two models is 1) for 5% significance level. This procedure was repeated for 1000 simulations.

A count to reject the null hypothesis is made for each statistics every time the statistics are greater than the specified chi-squared critical value. The estimate of the power of each test is calculated as the sum of these counts divided by 1000 (the number of simulations) for each value of h. I created a plot having both test statistics, h on the x-axis against the power of the test on the y-axis.

In the graphs, the score statistics and the log-likelihood ratio test produced similar, but not identical results. It can also be seen that the power of the tests values depends on the sample size.



δ=-2.1,N=8,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

δ=-2.1,N=5,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

δ=-1,N=10,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

δ=1,N=8,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

δ=-1,N=5,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

δ=1,N=5,$\lambda_2 = \lambda_1 * h$,$\lambda_1$=6

Figure 6.1: Simulations with small sample size

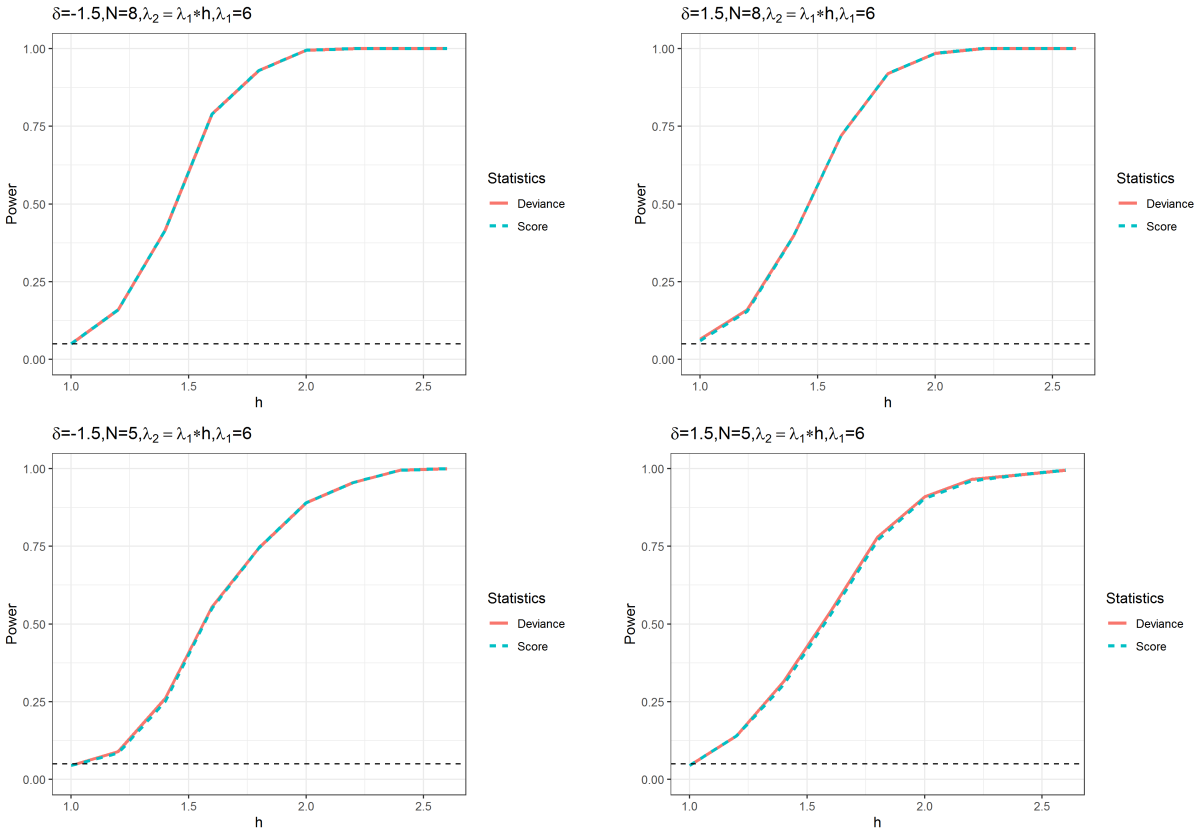


Figure 6.2: Simulations with small sample size

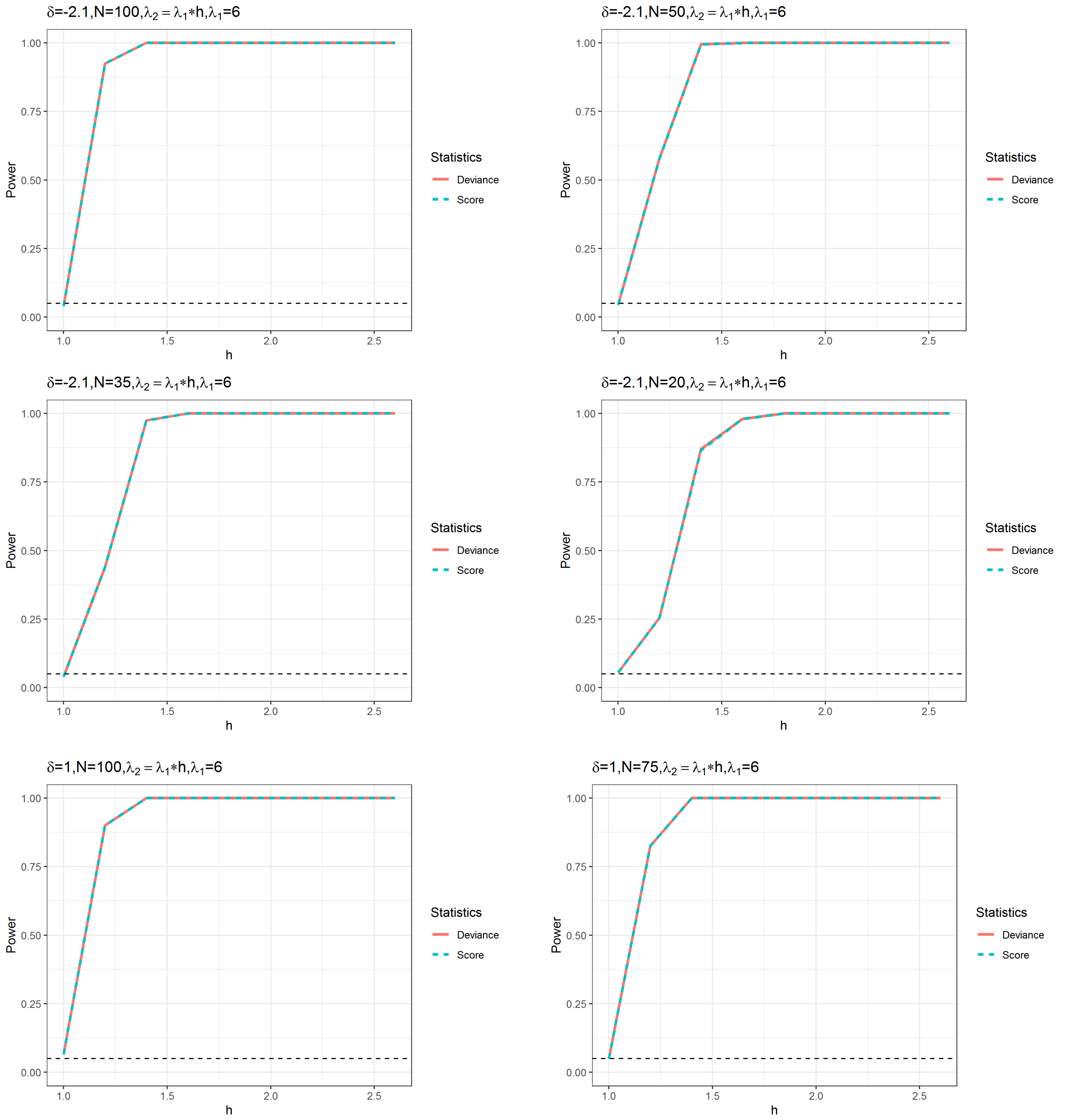


Figure 6.3: Simulations with large sample size

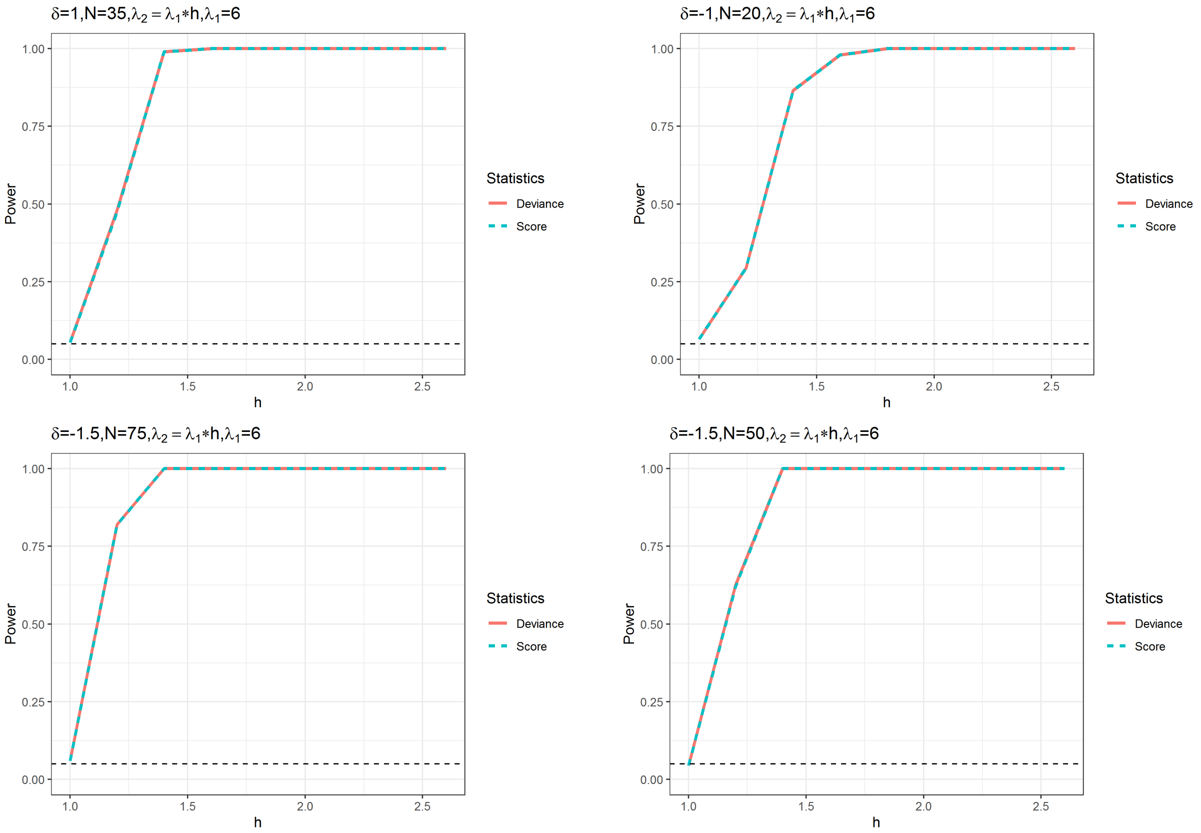


Figure 6.4: Simulations with large sample size

7. CONCLUSION

In this thesis, a generalization of the Poisson distribution for count data, the Touchard distribution, was studied. The main objective of most generalizations of the Poisson model is the extent to which it fits count data, that is, the extent to which it curtails overdispersion, underdispersion, and the presence of excess zeros. This thesis is divided into seven chapters. In Chapter 2, we recalled some properties of the exponential family of distributions, in particular the method of maximum likelihood estimation.

Furthermore, in Chapter 3 we established that the Touchard distribution with known δ is a member of the exponential family of distribution and thus has the properties of the exponential family of distributions. Also in this chapter, we established the analysis of the Touchard model of zero-inflated count, and also compared it with the Poisson model by a ratio of their probabilities of zeros. Important properties such as the index of dispersion, skewness and kurtosis, and the maximum likelihood estimation of the parameters of the Touchard distribution were also discussed.

Chapters 4 and 5 were devoted to the generalized linear model and its inferential statistics. Here, emphasis was placed on the log-likelihood ratio test and the score statistics. Chapter 5 presented the comparison of two samples with Touchard distribution, parameter estimation by MLE and hypothesis testing for this Touchard model. The simulations were performed in R software [version 3.6.0, https://cran.r-project.org/], and the results showed that it is recommended to apply the Touchard model to data with large sample size. Also, differences in the performance between the tests based on likelihood ratio test or score statistics were not observed.

Bibliography

[1] Dobson, A. J., Barnett, A. G. Introduction to Generalized Linear Models, 3rd. ed., Chapman and Hall, 2008.

[2] Raul Matsushita, Donald Pianto, Bernardo B. De Andrade, Andre Cançado & Sergio Da Silva: The Touchard distribution, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2018.1444177, 2018.

[3] McCullagh, P. and Nelder, J. A.. Generalized Linear Models, volume 37 of Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2 edition, 1989.

[4] Hogg, Robert V., Joseph W. McKean, and Allen T. Craig. Introduction to Mathematical Statistics. 7th ed. Boston: Pearson, 2013.

[5] Johnson, N. L., A.W. Kemp, and S. Kotz. Univariate discrete distributions. New York:Wiley. 2005.

[6] Haight, Frank A., Handbook of the Poisson Distribution, New York, NY, USA: John Wiley & Sons, ISBN 978-0-471-33932-8, 1967.

[7] Touchard, J., Sur les cycles des substitutions. Acta Math. 70:243–79, 1939.

[8] Chrysaphinou, O. On Touchard polynomials. Discrete Math. 54:143–52, 1985.

[9] Rota, G.-C. The number of partitions of a set. Amer. Math. Monthly 71:498–504, 1964.

[10] John Aldrich. ”R.A. Fisher and the making of maximum likelihood 1912-1922.” Statist. Sci. 12 (3) 162 - 176, https://doi.org/10.1214/ss/1030037906, August 1997.

[11] Lambert, Diane. ”Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing”. Technometrics. 34 (1): 1–14. doi:10.2307/1269547. JSTOR 1269547, 1992.

[12] Pesta, Michal. Poisson regression [online],. Available from: https://www2.karlin.mff.cuni.cz/ pesta/NMFM404/poisson.html, [cit. 04/29/2021]

[13] Songfeng Andy Zheng. Sufficient Statistics and Exponential Family [online], Available from: https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Sufficient.pdf, [cit. 04/30/2021].

[14] Gilthorpe MS, Frydenberg M, Cheng Y, Baelum V. Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. Stat Med. 2009 Dec 10;28(28):3539-53. doi: 10.1002/sim.3699. PMID: 19902494, 2009.

A. Powers of the Tests

```
library(touchard)
library(rootSolve)
library(ggplot2)
library(latex2exp)

csq = qchisq(0.95,1) #chisqaure quantile
n_sim = 1000 #Number of simulations
h_vec = seq(1,2.6,by=0.2) #values for h

#Values of N and Delta
N_vec <- c(5, 8, 10, 14, 16)  #c(20,35,50,75, 100)
delta_vec = c(-1.5, -1, 1, 1.5)

#solves for lambda for given mu
solve_lambda <- function(mu_in, delta) {
  f <- function(lbd)(tau(lbd,delta+1)/tau(lbd, delta))-1- mu_in
  ld <- uniroot.all(f, c(0,100))
  return(ld)
}

#solves for variance
t_var <- function(l_in,d_in) {
  v_out <-(tau(l_in,d_in+2)/tau(l_in,d_in)) -(tau(l_in,d_in+1)/tau(l_in,d_in))*(tau(l_in,d_in+1)
  /tau(l_in,d_in))
  return(v_out)
}


for ( N in N_vec) {
  for ( delta in delta_vec) {
    #############
    r_rej_vec_dev = c()
    r_rej_vec_sco = c()

    for ( h in h_vec) { #For loop for each h

      lambda_1 <- 6
      lambda_2 <- lambda_1 * h
      nrej_dev = 0
      nrej_sco = 0

      for ( i in seq(1,n_sim,by=1)) { #The 1000 Simulations for loop

        #The two random samples
        Y_1 <- rtouch(n = N, lambda=lambda_1, delta=delta)
        Y_2 <- rtouch(n = N, lambda=lambda_2, delta=delta)
        Y <-data.frame(Y_1,Y_2)

        #######Hypothesis############
        #Null Hypothesis
        mu_hat = (1/(2*N))*sum(Y)
        ld <- solve_lambda(mu_hat, delta)

      #Alternative Hypothesis
      mu_hat2 = (1/N)*sum(Y$Y_2)
      mu_hat1 = ((1/N)*sum(Y)) - mu_hat2

      ld_1 <- solve_lambda(mu_hat1, delta)
      ld_2 <- solve_lambda(mu_hat2, delta)

      #############Score######################################
      ######score######
      U = c(0, sum(Y$Y_2)-(N*mu_hat))
      J_inv = solve( t_var(ld, delta) * rbind(c(2*N,N),c(N,N)) )
      score = t(U)%*% J_inv %*% U

      #############Deviance######################################
      ##DElta D###
      y = c(as.matrix(Y))
      del_D = 2 *(sum(log( dtouch(y, lambda = c(rep(ld_1,N),
                  rep(ld_2,N)), delta = delta)))
                  -
                    sum(log( dtouch(y, lambda =ld,      delta = delta)))
      )


      # Verify if delta_D is greater than the chiSquare
      if(del_D  > csq ){
        nrej_dev = nrej_dev + 1
      } else {
        #do nothing
      }

      # Verify if score is greater than the chiSquare
      if(score  > csq ){
        nrej_sco = nrej_sco + 1
      } else {
        #do nothing
      }
    }

    r_rej_dev = nrej_dev / n_sim
    r_rej_vec_dev = c(r_rej_vec_dev, r_rej_dev)

    r_rej_sco = nrej_sco / n_sim
    r_rej_vec_sco = c(r_rej_vec_sco, r_rej_sco)

  }

  #Plot data
  df <- data.frame(Statistics = c(rep("Score", each=length((h_vec))),
                 rep("Deviance", each=length((h_vec)))),
                  h=c(h_vec,h_vec),
                  Power=c( r_rej_vec_sco , r_rej_vec_dev) )
  #Plot title and legend
  tt= paste0("$\\delta =",delta," , N =",N," , \\lambda_{2} =
            \\lambda_{1} * h "," , \\lambda_{1} =",lambda_1,"$")
  img_plot <- ggplot(df, aes(x=h, y=Power, group=Statistics)) +
    geom_line(aes(linetype=Statistics, color=Statistics), size=1.2)+
      coord_cartesian(ylim=c(0, 1)) +
      geom_hline(yintercept=0.05, linetype=”dashed”) +
      theme_bw() +
      ggtitle(TeX(tt))
    #Saves plots to file for each plot
    ggsave(img_plot, file=paste0(”plot_del”, delta,”N”,N,”.png”),
                    width = 15.61, height = 11.22, units = ”cm”)


  }
}
```

B. R: Variance-Mean Dependency

```r
##### Variance- Mean Dependency Plot
library(touchard)
library(ggplot2)
library(latex2exp)


r <- matrix(NA, nrow=7, ncol=11)
lambda <- seq(0.05,100,by=0.1) #values of lambda
delta <- c(-7,seq(-5,10,by=5)) #selected values of delta

datalist = list()

#calculate mu and variance for each values of lambda and delta
for ( j in delta) {
  ex = c()
  va = c()
  for ( i in lambda ) {
    t0 <- tau(i,j)
    t1 <- tau(i,j+1)
    t2 <- tau(i,j+2)
    ex <- c(ex,(t1/t0) - 1)
    va <- c(va,(t2/t0) -(t1/t0)*(t1/t0))
  }
  dat <- data.frame(ex, va)
  dat$j <- j
  datalist[[which(j==delta)]] <- dat # add it to your list
}

big_data = do.call(rbind, datalist)
#plot the data
p <- ggplot(big_data, aes(x=ex, y=va, color=as.factor(j)))+
  geom_line(size=0.8) +
  xlab(TeX('$\\mu$')) +
  ylab(TeX('$\\sigma^2$')) +
  ggtitle(TeX(' ')) +
  coord_cartesian(ylim=c(-1, 70),xlim=c(-1, 50)) +
  guides(color=guide_legend(title=NULL)) +
  scale_color_discrete(labels=lapply(
              sprintf('$\\delta = %d$', delta), TeX)) +
  theme_bw()
print(p)
```

C. R: Kurtosis

```
library(touchard)
library(latex2exp)


lambdas = seq(1, 20, by=0.1) #values of lambda
deltas =  c(-1,-1.5,-2)   # c(1,1.5,2) c(-0.5,0,0.5)values of delta

lines <- matrix(NA, nrow=length(lambdas), ncol=3)
#store values in a matrix

for (del in deltas) {

  e_r_vec = c() #vec of values for each delta
  for (lambda in lambdas) {
    t0 <- tau(lambda,del) #Tau
    t1 <- tau(lambda,del+1)
    t2 <- tau(lambda,del+2)
    ex <- (t1/t0) - 1 #mu
    va <- (t2/t0) -(t1/t0)*(t1/t0) #variance
    #Fourth moment
    r = 4
    e_r_fth = 0
    for (j in seq(0,r, by=1)) {
      e_r_fth = e_r_fth +
        choose(r, j)*(
          ((-1)^(r-j))*tau(lambda=lambda,
                delta=(del+j)) / tau(lambda=lambda, delta=del)
        )
    }
    #Third moment
    r = 3
    e_rthird = 0
    for (j in seq(0, r ,by=1)) {
      e_rthird = e_rthird +
        choose(r, j)*(
          ((-1)^(r-j))*tau(lambda=lambda, delta=(del+j)) /
                    tau(lambda=lambda, delta=del)
        )
    }

    kurt = (e_r_fth -4*e_rthird*ex +6*( va +ex^2)*ex*ex - 3* ex^4) / va^2
            #Index of kurtosis
    e_r_vec = c(e_r_vec, kurt)
  }
  lines[,which(del == deltas)] = e_r_vec
}

t1= paste0("$\\delta =",deltas[1],"$") #Plot legends
t2= paste0("$\\delta =",deltas[2],"$")
t3= paste0("$\\delta =",deltas[3],"$")
y_lab = paste0("$ E(X^",r,")$")
#Plot data
plot(lambdas, lines[,1], type = "l", lwd=2, font.axis = 2,
     ylab = 'Kurtosis', xlab = TeX("$\\lambda$"), col = rgb(0,0,0))  #, ylim = c(0,200)
lines(lambdas, lines[,2], type = "l", col = rgb(1,0,0), lwd=2)
lines(lambdas, lines[,3], type = "l", col = rgb(0,1,0), lwd=2)
legend("topright",  bty="n",                                      # Add legend to plot
       legend = c(TeX(t1), TeX(t2), TeX(t3)),
       col = c(rgb(0,0,0),rgb(1,0,0),rgb(0,1,0)),
       pch = c(16, 16, 16))
```

D. R: Ratio of the probabilities of zeros

```r
#Ratio of Probabilities of zeros
library(touchard)
library(latex2exp)

lambda = seq(1, 20, by=0.01)
delta =c(1,1.5,2)# c(-0.5,0,0.5) c(-1,-1.5,-2)

p_zero1 = dtouch(0, lambda=lambda, delta=delta[1])
                                / dpois(0, lambda=lambda)
p_zero2 = dtouch(0, lambda=lambda, delta=delta[2])
                               / dpois(0, lambda=lambda)
p_zero3 = dtouch(0, lambda=lambda, delta=delta[3])
                               / dpois(0, lambda=lambda)

t1= paste0("$\\delta =",delta[1],"$")
t2= paste0("$\\delta =",delta[2],"$")
t3= paste0("$\\delta =",delta[3],"$")

plot(lambda, p_zero1, type = "l", lwd=2, font.axis = 2,
     ylab = "P(0) Ratio", xlab = "Lambda", col = rgb(0,0,0),
                 ylim = c(0,0.8))    #, ylim = c(0,200) # Draw first line
lines(lambda, p_zero2, type = "l", col = rgb(0,0,1), lwd=2)
                 # Add second line
lines(lambda, p_zero3, type = "l", col = rgb(0,1,0), lwd=2)
legend("topleft", bty="n",
                           # Add legend to plot
       legend = c(TeX(t1), TeX(t2), TeX(t3)),
       col = c(rgb(0,0,0),rgb(0,0,1),rgb(0,1,0)),
       pch = c(16, 16, 16))
```

E. R: Skewness

```
# Index of Skewness
library(touchard)
library(latex2exp)

r= 3
lambdas = seq(1, 20, by=0.1)
deltas =   c(1,1.5,2)   #c(-0.5,0,0.5)  c(-1,-1.5,-2)

lines <- matrix(NA, nrow=length(lambdas), ncol=3)

for (del in deltas) {
  e_r_vec = c() #vec of values for each delta
  for (lambda in lambdas) {

    t0 <- tau(lambda,del)
    t1 <- tau(lambda,del+1)
    t2 <- tau(lambda,del+2)
    ex <- (t1/t0) - 1
    va <- (t2/t0) -(t1/t0)*(t1/t0)

    e_r = 0
    for (j in seq(0,r,by=1)) {
      e_r = e_r +
            choose(r, j)*(
              ((-1)^(r-j))*tau(lambda=lambda, delta=(del+j))
                        / tau(lambda=lambda, delta=del)
                          )
    }
    skew = ( e_r -3*ex*va - ex^3)/ va^(3/2)
    e_r_vec = c(e_r_vec, skew)
  }
  lines[,which(del == deltas)] = e_r_vec
}

t1= paste0("$\\delta =",deltas[1],"$")
t2= paste0("$\\delta =",deltas[2],"$")
t3= paste0("$\\delta =",deltas[3],"$")
y_lab = paste0("$ E(X^",r,")$")

plot(lambdas, lines[,1], type = "l", lwd=2, font.axis = 2,
     ylab = 'Skewness', xlab = TeX("$\\lambda$"),
                col = rgb(0,0,0), ylim = c(0.2,0.9))
lines(lambdas, lines[,2], type = "l", col = rgb(1,0,0), lwd=2)
lines(lambdas, lines[,3], type = "l", col = rgb(0,1,0), lwd=2)
legend("topleft", bty="n",
       legend = c(TeX(t1), TeX(t2), TeX(t3)),
       col = c(rgb(0,0,0),rgb(1,0,0),rgb(0,1,0)),
       pch = c(16, 16, 16))
```