NOVEL REGRESSION MODELS FOR DISCRETE RESPONSE

Author:
Alina Peluso

Supervisor:
Veronica Vinciotti

A thesis submitted for the degree of
Doctor of Philosophy

Department of Mathematics

December 2017

Abstract

NOVEL REGRESSION MODELS FOR DISCRETE RESPONSE

In a regression context, the aim is to analyse a response variable of interest conditional to a set of covariates. In many applications the response variable is discrete. Examples include the event of surviving a heart attack, the number of hospitalisation days, the number of times that individuals benefit of a health service, and so on. This thesis advances the methodology and the application of regression models with discrete response. First, we present a difference-in-differences approach to model a binary response in a health policy evaluation framework. In particular, generalized linear mixed methods are employed to model multiple dependent outcomes in order to quantify the effect of an adopted pay-for-performance program while accounting for the heterogeneity of the data at the multiple nested levels. The results show how the policy had a positive effect on the hospitals' quality in terms of those outcomes that can be more influenced by a managerial activity. Next, we focus on regression models for count response variables. In a parametric framework, Poisson regression is the simplest model for count data though it is often found not adequate in real applications, particularly in the presence of excessive zeros and in the case of dispersion, i.e. when the conditional mean is different to the conditional variance. Negative Binomial regression is the standard model for over-dispersed data, but it fails in the presence of under-dispersion. Poisson-Inverse Gaussian regression can be used in the case of over-dispersed data, Generalised-Poisson regression can be employed in the case of under-dispersed data, and Conway-Maxwell Poisson regression can be employed in both cases of over- or under-dispersed data, though the interpretability of these models is not straightforward and they are often found computationally demanding. While Jittering is the default non-parametric approach for count data, inference has to be made for each individual quantile, separate quantiles may cross and the underlying uniform random sampling can generate instability in the estimation. These features motivate the development of a novel parametric regression model for counts via a Discrete Weibull distribution. This distribution is able to adapt to different types of dispersion relative to Poisson, and it also has the advantage of having a closed form expression for the quantiles. As well as the standard regression model, generalized linear mixed models and generalized additive models are presented via this distribution. Simulated and real data applications with different type of dispersion show a good performance of Discrete Weibull-based regression models compared with existing regression approaches for count data.

Acknowledgements

I gratefully acknowledge the research centre CRISP (Centro di Ricerca Inter-universitario per i Servizi di Pubblica utilitá - University of Milan-Bicocca) for providing the health data of Lombardy region of Italy which have been used throughout the thesis. Besides, I would like to express my appreciation to many people who have been contributed their time, expertise and talent to complete my study and making my time at Brunel University London successful. I am very proud and deeply grateful to my supervisor Dr. Veronica Vinciotti for her constant and highly valuable support, help, advice and her supervision of this thesis. I also want to express my sincere gratitude to my colleague Paolo Berta for being supportive and always offering great advices. I also would like to thank my second supervisor Dr. Francesco Moscone. My thanks also go to my colleague and dear friend Farukh Mukhamedov. Finally, I extremely appreciate the support, help and love of my mum and my sister.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Modelling discrete response

The main assumption in Ordinary Least Squares regression (OLS) is that the dependent variable is continuous. There are numerous real world processes whose outcomes are count variables, e.g. the number of days spent in hospital, the number of deaths recorded for a specific condition, the number of doctor visits, and so on. In some cases, the dependent variable takes the value of zero for many observations, e.g. the number of patients affected by a rare health condition. In other cases, the dependent variable is binary, e.g. an event which either did or did not occur such as being exposed to a treatment, or the event of surviving a disease. This thesis describes alternatives and extensions to existing regression methods for these types of data. In particular the focus will be on logistic regression and on regression models for count data. This chapter will introduce these topics and will conclude with the contribution of the thesis to this field.

1.2 Generalized linear and mixed regression models

1.2.1 Generalized linear model

Generalized linear models (GLMs) [67, 72] relax the assumptions made by linear regression models that the response variable is continuous and normally distributed conditional on the predictors. Let Y be the response variable and $X = (X_1, \ldots, X_P)^T$ the vector of P predictors. The conditional distribution of $Y|X$ is assumed to belong to an exponential family, and it has probability function

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

where λ is the canonical parameter while ϕ is the dispersion parameter. The functions $a(\cdot)$, and $c(\cdot)$ are known and determine the type of distribution. The parameters λ and ϕ can be also defined as location and scale parameters, respectively. Thus, when the response variable has a distribution in the exponential family, its conditional mean can be written as $E(Y|X) = \mu = a'(\lambda)$ and its conditional variance is in the form of $\text{var}(Y|X) = \phi a''(\lambda)$, where $a'(\lambda)$ and $a''(\lambda)$ are the first and second derivative of $a(\lambda)$. This means that, up to a dispersion parameter ϕ, the distribution of Y is determined by its mean. Moreover, the dependence of the conditional mean on the regressors is specified as

$$g(\mu(x)) = x\theta,$$

where $g(\cdot)$ is a known link function, $x = (1, x_1, \ldots, x_P)$, and $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$ is the vector of regression coefficients. The link function can take different forms. In particular, the canonical link is defined when the link function makes the linear predictor equal to the canonical parameter $\mu(x)$, as in the case of a standard linear regression, while the logit is the respective link for the Binomial distribution, and the log is the link for both the Poisson and the Negative Binomial distribution. These models are described in detail below. The parameters of these models are typically estimated by maximum likelihood.

Logistic model Many count response variables are binary, that is the response variable can take two possible outcomes only. In this case, the distribution of the response is specified by the probability $P(Y = 1) = \pi$ of success, and by the probability $P(Y = 0) = (1 - \pi)$ of failure. Thus, the conditional distribution is given by

$$Y|X \sim \text{Binomial}(n, \pi(x)),$$

where the probability function is defined by

$$f(y; n, \pi(x)) = \binom{n}{y} (\pi(x))^y (1 - \pi(x))^{(n-y)},$$

with $0 < \pi(x) < 1$. The logit link is typically used to link π with x, i.e.

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = x\theta,$$

from which $\pi(x) = \frac{\exp(x\theta)}{1+\exp(x\theta)}$ gives the popular sigmoid relationship which guarantees that $\pi(x)$ is between 0 and 1 for any real values of θ.

Poisson model The Poisson regression is the simplest count model upon which a variety of other count models are based on. In this case,

$$Y|X \sim \text{Poisson}(\mu(x)),$$

that is

$$f(y; \mu(x)) = \frac{e^{-\mu(x)}(\mu(x))^y}{y!}$$

for $y = 0, 1, 2, \ldots$, and $\mu(x) > 0$. Here μ is linked to the predictors x via

$$\log(\mu(x)) = x\theta,$$

that is there is a log-linear relationship between the mean and the predictors. For the properties of a Poisson distribution, $E(Y|X) = \text{var}(Y|X) = \mu(x)$ making this regression model too restrictive in many applications.

Negative Binomial model Unlike the Poisson distribution, the variance of a Negative Binomial differs from its mean. The Negative Binomial distribution can be defined as

$$Y|X \sim \text{Negative Binomial}(\mu(x), \sigma),$$

and its probability function assumes the form

$$f(y; \mu(x), \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left(\frac{\sigma\mu(x)}{1 + \sigma\mu(x)}\right)^y \left(\frac{1}{1 + \sigma\mu(x)}\right)^{\frac{1}{\sigma}}$$

for $y = 0, 1, 2, \ldots$, $\mu(x) > 0$, and $\sigma > 0$. This parametrization is equivalent to that used by [8] except that there $\alpha = \frac{1}{\sigma}$ instead of σ. For this parametrisation, $\mu(x)$ is the conditional mean, and $\mu(x) + \sigma(\mu(x))^2$ is the conditional variance. The σ parameter is referred to as the dispersion parameter. Since $\mu(x) + (\mu(x))^2\sigma \geq 0$, this model can only account for over-dispersion relative to Poisson. In a Negative Binomial model, the mean is linked to the predictors via

$$\log(\mu(x)) = x\theta.$$

Poisson-inverse Gaussian model The Poisson-inverse Gaussian distribution was first introduced by [51]. This distribution is a two parameter mixture of the Poisson distribution and the inverse Gaussian distribution. Due to the flexibility of the inverse Gaussian distribution, the Poisson-inverse Gaussian distribution is particularly useful for modelling over-dispersed count data. In particular [111] proposed this distribution as an alternative to the Negative Binomial. The probability function of the Poisson-inverse Gaussian distribution is given by

$$Y|X \sim \text{Poisson-inverse Gaussian}(\mu(x), \sigma),$$

where

$$f(y; \mu(x), \sigma) = \left(\frac{2\alpha^{\frac{1}{2}}}{\pi}\right) \frac{\mu(x)^y e^{\frac{1}{\sigma}} K_{y-\frac{1}{2}}(\alpha)}{(\alpha\sigma)^y y!},$$

for $y = 0, 1, 2, \ldots$, with $\mu(x) > 0$, $\sigma > 0$, $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu(x)}{\sigma}$, and where $K(\cdot)$ is a modified Bessel function of the third kind (see Chapter 10 in [2]). This parametrization is used in [30]. The Poisson-inverse Gaussian distribution can be seen as a special case of the Sichel$(\mu(x), \sigma, \nu)$, when $\nu = -\frac{1}{2}$. For the Poisson-inverse Gaussian distribution $\mu(x) + (\mu(x))^2\sigma$ is the conditional variance, and $\mu(x)$ is the conditional mean, which is linked to the predictors as follows

$$\log(\mu(x)) = x\theta.$$

Conway–Maxwell–Poisson model It has been shown above how Poisson and Negative Binomial models can only account for equi- and over-dispersed data, respectively. Other models have been developed that can account for different types of dispersion. Among others, the Conway–Maxwell–Poisson (COM-Poisson) distribution is a generalisation of the Poisson distribution which allows to model both under-dispersed and over-dispersed data. One of the properties of the Poisson distribution is that the ratio of consecutive probabilities is linear in y, i.e. $\frac{P(Y=y-1|X)}{P(Y=y|X)} = \frac{y}{\mu(x)}$, as showed in [68]. However, in some applications this ratio may not decrease linearly in y, i.e. the distribution may have a thinner or thicker tail than the Poisson. Generalising the above formulation leads to the ratio $\frac{y^\sigma}{\mu(x)}$, and the COM-Poisson is the distribution for which this holds. In particular, this distribution can be described as

$$Y|X \sim \text{COM-Poisson}(\mu(x), \sigma)$$

which has probability function

$$f(y; \mu(x), \sigma) = \frac{(\mu(x))^y}{(y!)^\sigma} \frac{1}{Z(\mu(x), \sigma)},$$

for $y = 0, 1, 2, \ldots$, $\mu(x) > 0$, $\sigma \geq 0$, and where the function $Z(\mu(x), \sigma) = \sum_{j=0}^{\infty} \frac{(\mu(x))^j}{(j!)^\sigma}$ serves as a normalisation constant so that the probability mass function sums to 1. When $\sigma = 1$, $Z(\mu(x), \sigma) = e^{\mu(x)}$, so the COM-Poisson distribution equals the formulation of the Poisson distribution. For the COM-Poisson distribution, $E(Y|X) = \sum_{j=0}^{\infty} \frac{j(\mu(x))^j}{(j!)^\sigma Z(\mu(x),\sigma)}$ is the conditional mean, and $\text{var}(Y|X) = \sum_{j=0}^{\infty} \frac{j^2(\mu(x))^j}{(j!)^\sigma Z(\mu(x),\sigma)}$ is the conditional variance. These and other moments cannot be computed in closed form, leading to computational issues when performing parameter inference using this distribution. In a COM-Poisson

regression model, the parameter μ is linked to the predictors via

$$\log(\mu(x)) = x\theta.$$

Generalised Poisson model Another distribution which has been proposed for modelling under-dispersion is the generalized Poisson distribution [26] which is a generalization of a Poisson distribution with an additional parameter being added. In particular, this distribution can be described as

$$Y|X \sim \text{Generalised Poisson}(\mu(x), \sigma)$$

which has probability function

$$f(y;\mu(x),\sigma) = \mu(x)(\mu(x)+\sigma y)^{y-1}\frac{\exp(-\mu(x)-\sigma y)}{y!},$$

for $y = 0,1,2,\ldots$, $\mu(x) > 0$, $\max(-1,-\frac{\theta}{m}) \leqslant \sigma \leqslant 1$, and where $m \geqslant 4$. The Poisson distribution corresponds to the case where σ=0. The weakness of the generalised Poisson model, however, is its inability to capture some levels of dispersion because the distribution is truncated under certain conditions on the dispersion parameter. For the generalised Poisson distribution, $E(Y|X) = \frac{\mu(x)}{1-\sigma}$ is the conditional mean, and $\text{var}(Y|X) = \frac{\mu(x)}{(1-\sigma)^3}$ is the conditional variance. The parameter μ is linked to the predictors via

$$\log(\mu(x)) = x\theta.$$

1.2.2 Excess zeros model

Zero inflated model

Zero inflated models are employed in the presence of an excess of zero counts in the response. As detailed in [18], these models are two-component mixture models combining zeros coming from both a point mass at zero and a conditional count distribution, i.e. $f(Y = y|X)$ (or shortly $f(y)$). The zeros are modelled through a binomial model, typically with logit or probit link. The zero realisations are modelled with probability $\pi(x)$, while the non-zeros with probability $(1-\pi(x))$. Thus,

$$Pr(Y|X) = \begin{cases} \pi(x) + (1-\pi(x))f(0) & \text{for } y = 0 \\ (1-\pi(x))f(y) & \text{for } y = 1,2,3,\ldots, \end{cases}$$

where $0 < \pi(x) < 1$ is the mixture proportion. Specifically, the mixture parameter can take any link function which maps it into $(-\infty,+\infty)$. The logit link is the preferred choice, thus π can be related to the set of covariates as

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x\gamma,$$

which leads to $\pi(x) = \frac{\exp(x\gamma)}{1+\exp(x\gamma)}$. For the second part of the mixture, i.e. $f(y)$, one can use any of the models described before.

Zero inflated Poisson model The zero inflated Poisson regression model which links both π and the mean to the predictors can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) + (1-\pi(x))e^{-\mu(x)} & \text{for } y = 0 \\ (1-\pi(x))\frac{e^{-\mu(x)}(\mu(x))^y}{y!} & \text{for } y = 1,2,3,\ldots. \end{cases} \tag{1.1}$$

Zero inflated Negative Binomial model The zero inflated Negative Binomial regression model which links both π and the mean to the predictors can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) + (1-\pi(x))\left(\frac{1}{1+\sigma\mu(x)}\right)^{\frac{1}{\sigma}} & \text{for } y = 0 \\ (1-\pi(x))\frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)}\left(\frac{\sigma\mu(x)}{1+\sigma\mu(x)}\right)^{y}\left(\frac{1}{1+\sigma\mu(x)}\right)^{\frac{1}{\sigma}} & \text{for } y = 1,2,3,\ldots. \end{cases} \tag{1.2}$$

Hurdle model

Hurdle models were first discussed by [27], but [70] significantly contributed to their application in modelling count data. Unlike the zero-inflated models, here the main idea is to partition the model estimating a process for the zero counts, and another for the positive counts via a zero-truncated count model, i.e. Poisson, Negative Binomial, or any other count distribution of interest. Thus, assuming that the zero counts are generated by a binary process $\pi(x)$, and the positive counts by a zero-truncation of a density $f(y)$, it follows that

$$Pr(Y|X) = \begin{cases} \pi(x) & \text{for } y = 0 \\ (1-\pi(x))\frac{f(y)}{1-f(0)} & \text{for } y = 1,2,3,\ldots. \end{cases}$$

Hurdle Poisson model The hurdle Poisson regression model which links both π and the mean to the predictors can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) & \text{for } y = 0 \\ (1-\pi(x))\frac{(\mu(x))^{y}e^{-\mu(x)}}{(1-e^{-\mu(x)})y!} & \text{for } y = 1,2,3,\ldots. \end{cases} \tag{1.3}$$

Hurdle Negative Binomial model The hurdle Negative Binomial regression model which links both π and the mean to the predictors can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) & \text{for } y = 0 \\ (1-\pi(x))\frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)}\left(\frac{\sigma\mu(x)}{1+\sigma\mu(x)}\right)^{y}\frac{1}{(1+\sigma\mu(x))^{\frac{1}{\sigma}}-1} & \text{for } y = 1,2,3,\ldots. \end{cases} \tag{1.4}$$

1.2.3 Generalized linear mixed model

Standard regression models assume that the observations are independent of each other conditional on X. This is not appropriate in the case of correlated data structures, specifically for clustered or longitudinal data. In these studies, subjects are observed nested within larger units, e.g. hospitals, countries, or repeated observations within subjects. Data with this structure are often referred to as multilevel or hierarchical data because the level-1 observations, i.e. subjects, are nested within the higher level-2 observations, i.e. clusters. Higher levels are also possible, e.g. a three-level study could have repeated level-1 observations nested within level-2 subjects which are nested within level-3 groups. Regression models for the analysis of such multilevel data are referred to as generalized linear mixed models (GLMMs). These are an extension to the GLM in which the linear predictor contains random effects in addition to the usual fixed effects. The basic idea underlying a random effects model is that the heterogeneity across individuals in the regression coefficients can be represented by additional random variables. In particular, the expected value of the outcome is related to the linear predictors through the link function

$$g(\mu(x,u)) = x\theta + zu,$$

where $x = (1, \ldots, x_p)$, $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$, and $z = (z_1, \ldots, z_Q)$ is the $(1 \times Q)$ design vector for the $(Q \times 1)$ random effects $u = (u_1, \ldots, u_Q)^T$, which are assumed i.i.d as $u \sim \text{Normal}(0, \sigma_u^2)$. The variable σ_u^2 indicates the degree of heterogeneity of the subjects.

Logistic mixed model A logistic regression mixed effects model can be written as

$$\text{logit}(\pi(x,u)) = \log\left(\frac{\pi(x,u)}{1-\pi(x,u)}\right) = x\theta + zu,$$

from which $\pi(x,u) = \frac{\exp(x\theta+zu)}{1+\exp(x\theta+zu)}$.

Poisson mixed model A Poisson model which links the mean to the predictors via both fixed and random effects can be written as

$$\log(\mu(x,u)) = x\theta + zu.$$

Negative Binomial mixed model A Negative Binomial model which links the mean to the predictors via both fixed and random effects can be written as

$$\log(\mu(x,u)) = x\theta + zu.$$

1.3 Generalized additive models

1.3.1 Generalized additive model for location

Generalized additive models (GAMs) were first introduced by [96], and later they have been made popular by [47]. GAMs allows for rather flexible specifications of the dependence of the response on the covariates. Specifically, while the simplicity is retained through the additive form of the model, GAMs extend GLMs to non-linear covariate effects that may not be identified using traditional linear regression methods. In particular, smooth non-linear functions are applied to each individual predictors. As in the GLM framework, we consider a response variable which pertains to the exponential family distribution, and we assume $E(Y|X) = \mu(x)$. Thus, for a GAM model we can define the link

$$g(\mu(x)) = s_0 + \sum_{p=1}^{P} s_p(x_p),$$

where $s_p(\cdot)$ represents a generic smoothing function for the covariate x_p. The non-linear functions can be defined within specified families, such as polynomials. Like for the GLMs, different link functions can be used, such as a logit or a probit for binomial response, or a Poisson for count data, and so on. The parameter estimation is done through a combination of back-fitting and iteratively re-weighted least squares algorithm.

1.3.2 Generalized additive model for location, scale and shape

GLMs and GAMs can be extended to modelling all the parameters of a distribution. The resulting model are called generalized additive models for location, scale and shape (GAMLSS), and they were first introduced by [84]. In GAMLSS, the exponential family distribution assumption for the response variable $Y|X$ is relaxed and replaced by a general

distribution family D, i.e. $Y|X \sim D(Y|X,\theta)$, where $D \in \boldsymbol{D}$ can be any distribution with K distribution parameters, such as location (μ), scale (σ), and shape parameters. These parameters are linked to the covariates as follows

$$g_k(\mu_k(x)) = s_{k0} + \sum_{p=1}^{P} s_{kp}(x_p)$$

where $k = (1,\ldots,K)$, $s_{kp}(\cdot)$ is a generic smoothing function for the kth distributional parameter, and for the covariate x_p. Thus, all the parameters of the distribution can be modelled as smoothing functions of the explanatory variables, i.e. cubic splines [44], penalized splines [37], lowess [25], varying coefficient models [48], and so on. In particular, this approach facilitate the interpretation of all the parameters of the distribution which can be explicitly linked to the different moments of the distribution, i.e. mean, variance, skewness and kurtosis.

It is possible to extend the previous model to a more general formulation, which specifically focus on the inclusion of the random effects. Thus,

$$g_k(\mu_k(x,u)) = s_{k0} + \sum_{p=1}^{P} s_{kp}(x_p) + zu_k$$

where $u_k = ({u_1}^T,\ldots,{u_{Q_K}}^T)^T$ is the random effects vector of coefficients where each component is assumed to be distributed as Normal$(0,\sigma^2_{u_k})$. Thus, within this more general formulation, each parameter of the distribution can be modelled through a smoothing function of each of the explanatory variables x_p and of the random effects u.

Poisson GAMLSS random effects model The Poisson GAMLSS random effects regression model can be written as

$$\log(\mu(x,u)) = s_0 + \sum_{p=1}^{P} s_p(x_p) + zu.$$

Negative Binomial GAMLSS random effects model There can be situations where the assumption of a constant scale parameter is not appropriate. Thus, modelling the scale parameter as a function of the explanatory variables may be useful in explaining more variation of the data. As detailed in [94], modelling the dispersion parameter within the GLM framework was done by [73], [91] and [106]. Moreover, [82] introduced a class of additive models for mean and dispersion (MADAM), by including smooth functions for modelling simultaneously both μ and σ. A Negative Binomial GAMLSS random effects regression model can be written as

$$\log(\mu(x,u)) = s_{10} + \sum_{p=1}^{P} s_{1p}(x_p) + zu_1$$
$$\log(\sigma(x,u)) = s_{20} + \sum_{p=1}^{P} s_{2p}(x_p) + zu_2.$$

Maximum likelihood inference for these models is more complex and is typically done by iteratively estimating regression parameters for one link while keeping the other fixed (see Section 7.4 of [105]).

1.4 Non-parametric regression model for discrete response: the jittering method

At the other spectrum of parametric approaches for discrete response, there are numerous non-parametric quantile regression methods which focus on modelling individual quantiles of the distribution and link these to the predictors via a regression model, without making any assumption on the parametric form of the conditional distribution. Of particular notice for discrete responses are the quantile regression models for binary and multinomial response of [65] and [52], and the median regression approach with ordered response of [59]. For a general discrete response, the literature on quantile regression for counts is mainly dominated by the jittering approach of [64], which was also rephrased in a Bayesian framework by [58] in the context of an environmental epidemiology study. In these approaches, the fitted regression parameters are specific to the selected conditional quantile, by using quantile-specific loss functions. Performing inference across a range of quantiles provides a global picture of the conditional distribution of the response variable, without having to specify the parametric form of the conditional distribution. This has proven to be rather useful in practice, particularly in cases where the relationship between response and predictors is complex. In the case of a count, however, quantile regression analysis is complicated by the fact that a non-differentiable objective function is combined with a discrete dependent variable. In such a context, it is impossible to obtain valid asymptotic results for the distribution of the conditional quantiles using standard econometric tools. In particular, let Y be the response variable and $X = (X_1, \dots, X_P)^T$ the vector of P predictors. The main problem with the estimation of quantile regression when Y results from counts is that because Y has a discrete distribution, the conditional τ-quantile $\mu^{(\tau)}(Y|X)$ cannot be a continuous function of the parameter of interest. This limitation can be overcome by constructing a continuous random variable whose quantiles have a one to one relation with the quantiles of Y. In particular, to deal with this issue [64] suggests smoothing the data by introducing the jittering method. The basic idea of the jittering approach is to build a continuous r.v. whose quantiles have a known relationship with the quantiles of the response. This task is achieved by creating the auxiliary variable $Z = Y + U$, where $U \sim$Uniform(0,1) which has conditional τ-quantiles linked to the predictors x via

$$\log\left(\mu^{(\tau)}(Z|X)\right) = x\theta^{(\tau)},$$

where $x = (1, x_1, \dots, x_P)$, and $\theta^{(\tau)} = (\theta_0^{(\tau)}, \theta_1^{(\tau)}, \dots, \theta_P^{(\tau)})^T$ is the vector of regression coefficients for the τ-quantile. Thus, standard quantile regression can be applied to a monotonic transformation of $Z|X$. The monotonic transformation ensures that the estimated quantiles of Z are non-negative and that the transform quantile function is linear in the parameters. Specifically, how the covariates affect $\mu^{(\tau)}(Z|X)$ is of interest because

$$\mu^{(\tau)}(Y|X) = \lceil \mu^{(\tau)}(Z|X) - 1 \rceil,$$

where $\lceil \cdot \rceil$ represents the ceiling function. In other words, it is possible to recover $\mu^{(\tau)}(Y|X)$ on the basis of $\mu^{(\tau)}(Z|X)$. As described above, this method requires sampling from a Uniform distribution, thus [64] suggest averaging the quantile regression estimates across M jittered sample. The resulting estimator is more efficient than the one obtained from a single draw. However, these non-parametric approaches suffer from some drawbacks: inference has to be made for each individual quantile, separate quantiles may cross and, additionally in the case of the jittering method, the underlying uniform random sampling can generate instability in the estimation. For example, considering the waiting times data which we will analyse in chapter 4 of this thesis, Figure 1.1 shows the marginal relation between the response variable (y-axis) i.e. the waiting times days before intervention for the three health conditions CABG, PTCA and hip replacement in Lombardy region of Italy, and a continuous variable (x-axis) i.e. the age of the patient, under a jittering approach (top plot) while keeping all the other covariates used in the regression model fixed to their mean if continuous, and to their mode if discrete. For the same data and via the same model specification, the bottom plot shows the marginal relation between the response and the predictor obtained via the parametric Discrete Weibull regression model that is introduced in this thesis.

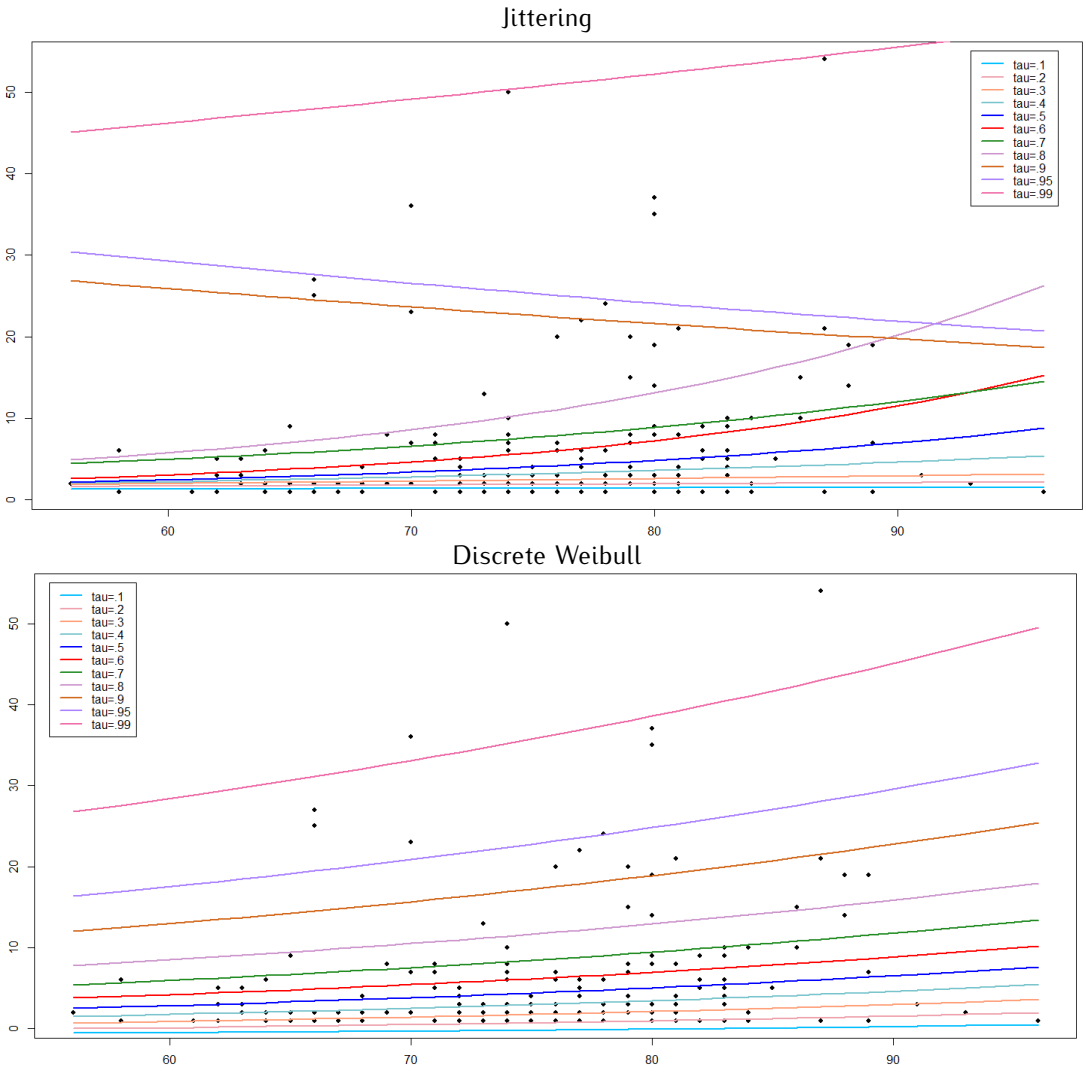


FIGURE 1.1: Plot of the τ-quantiles of the response (y-axis) over the variable age (x-axis) under a linear non-parametric (top) and parametric (bottom) fit of the model and while keeping all the other covariates fixed to their mean if continuous, and to their mode if discrete for the hospital waiting times data.

1.5 Contributions of the thesis

1.5.1 Difference-in-differences approach via a logistic linear mixed model

In chapter 2 we present a generalized linear mixed model for a binary response in a health policy evaluation framework through a difference-in-differences approach. In particular, the focus is on the evaluation of a pay-for-performance program, which is widely adopted to drive improvements in the quality of healthcare provision. In this field, previous studies evaluating the impact of these programs are either limited by the number of health outcomes or of medical conditions considered. Thus, our novel approach aims to evaluate the effectiveness of the adopted pay-for-performance program on the basis of five health outcomes, and across a wide range of medical conditions. The context of the study is Lombardy region of Italy, where a rewarding program was introduced in 2012. The model includes multiple dependent outcomes that allow quantifying the joint effect of the program, and random effects that account for the heterogeneity of the data at the ward and hospital level. Our results show that the policy had a positive effect on the hospitals' performances in terms of those outcomes that can be more influenced by a managerial activity, namely the number of readmissions, transfers

and returns to surgery room. No significant changes which can be related to the pay-for-performance introduction are observed for the number of voluntary discharges and for mortality. Moreover, our study shows evidence that the medical wards have reacted more strongly to the P4P program than the surgical ones, whereas only limited evidence is found in support of a different policy reaction across types of hospital ownership. Finally, the evaluation found no evidence of a distortion of the hospital behaviour aimed at inflating the performance evaluation such as cream skimming behaviour.

1.5.2 Linear models for counts via a Discrete Weibull distribution

In chapter 3 we focus on linear regression models for count data. This analysis was motivated by the lack of a unified and flexible regression framework for count response which can easily adapt to the different cases of dispersion, and in presence of excessive zeros in the data. This has been addressed with the use of a Discrete Weibull distribution. Beyond equi-dispersion, this model has the ability to capture over-dispersion, and under-dispersion relative to Poisson, i.e. all cases where the conditional variance is different to the mean. Moreover, this model is particularly flexible in the presence of an excess of zeros. In addition to the standard regression model, the analysis has been extended with the inclusion of multilevel models i.e. mixed effects models, which typically recognise the existence of hierarchies of the data measured at multiple nested levels. Simulated and real data examples are used to show the performance obtained via the Discrete Weibull model in comparison with existing parametric approaches.

1.5.3 Non-linear models for counts via a Discrete Weibull distribution

In chapter 4 we extend the regression framework for count response to more complex dependencies where linearity or the parametric form of the distribution may be too restrictive. Thus, we developed a Discrete Weibull-based generalized additive model. Different smoothing functions, i.e. polynomials and un-penalized regression spline of a variable, are proposed in the link function. Maximum likelihood is considered as well as a Lasso regression approach for variable selection. Lastly, Gaussian kernel weights are presented for a local regression approach. A comparison on simulated and real data studies is made with existing parametric approaches and non-parametric regression models for counts, such as the jittering method of [64]. We show how our flexible regression method can approximate well the conditional distribution of the response given the predictors across a number of quantiles, and how its performance is comparable to that of the non-parametric quantile regression approach of [64] which fits separate models for each conditional quantile.

Chapter 2

Difference-in-differences approach via a logistic linear mixed model

In this chapter we present a generalized linear mixed model for a binary response in a health policy evaluation framework through a Difference-In-Differences approach (DID) approach. In particular, the focus is on the evaluation of a pay-for-performance (P4P) program, which is widely adopted to drive improvements in the quality of healthcare provision. The analyses presented in this chapter have been performed in `SAS 9.3` software [54], and in `Stata 14` software [95].

2.1 Overview of the study

As part of a reforming project aimed to improve the efficiency, the effectiveness, and the quality of care delivered by a financially sustainable health system, in 2012 the Italian region of Lombardy adopted a P4P. Motivated by the necessity to evaluate the effectiveness of this newly introduced program, we investigate whether the P4P incentives have led to better health outcomes.

2.1.1 The P4P program and the experimental design for policy evaluation

Quality improvement is the principal strategy of any healthcare system. For this reason, there is a strong focus on assessment and redesign of the work process and of the systems themselves in order to lower the costs and to deliver care that is safer and which results in the best outcome for patients. The adoption of a P4P approach aims to drive the hospitals in this direction. The idea behind the implementation of a P4P approach is quite simple: in order to improve the overall quality delivered, healthcare providers are given the opportunity to have their reimbursement increased when they achieve specified quality benchmarks [7, 36]. From an economics perspective, the hospital is considered as a profit maximizer agent which is encouraged to compete for quality in order to obtain a financial reward, rather than to attract more patients. Therefore, a P4P program is considered efficient when an improved quality of care is achieved with equal or lower costs for the overall healthcare system [39]. Clearly the evaluation of the quality delivered is a crucial part to every P4P approach. While quality in healthcare is a broad concept composed of different dimensions, such as efficiency, appropriateness and customer satisfaction, P4P programs refer to the healthcare system's quality mostly in terms of its effectiveness [102].

Due to the potential of P4P programs, in recent years there has been a growing interest in the application of these programs to the healthcare systems of different countries. These studies are collected in several systematic reviews [35, 76, 103], but mixed results transpire about the impact of the programs to the quality of care. The aim of the current chapter is to contribute to the existing literature by providing a thorough evaluation of a P4P program and its effect on the overall quality of the healthcare system. The study discussed in this chapter pertains the Lombardy region of Italy, previously identified as a suitable context for the adoption of P4P program [23]. Data were collected both two years prior and two year post introduction of the policy for all hospitals in the Lombardy region. As data are available also two year post introduction of the policy, our analysis can reveal a possible delayed impact of the P4P program. In this way, we extend the existing literature with an evaluation of the impact beyond the immediate P4P introduction.

2.1.2 The specifics of the P4P program in Italy and its implementation

The Italian healthcare system provides universal healthcare coverage. The state government guarantees the Essential Levels of Assistance (LEA) over all regions of the country. Each region has administrative and executive freedom of implementation of the LEA, and citizens may freely choose the healthcare provider. The Italian NHS is funded mainly from general taxation. Financial resources for NHS are transferred from the state to a regional budget, and are then managed by the local healthcare system [66]. Among the 21 regions in Italy, Lombardy is one of the top-ranked for socio-demographic indicators and one of the most competitive areas in Europe according to economic indicators. Lombardy has a population of 10 million residents, equal to 16% of the total Italian population, with a density of 404 inhabitants per km^2. The Lombardy healthcare system comprises of circa 150 hospitals generating around 1.6 million discharges annually, with circa 18 billion Euro allocated for the healthcare spending i.e. circa 75% of the regional budget, every year.

A regional reform in 1997 radically transformed the healthcare system in Lombardy into a quasi-market healthcare system in which citizens can freely choose the provider regardless of its ownership (private for profit, private not for profit, or public). In particular, the healthcare system in Lombardy is entirely built on a prospective payment system based on a classification of the inpatient stay into groups for the purposes of payment, i.e. Diagnosis-related Groups (DRGs). The factors used to determine the DRGs payment amount include the diagnosis involved as well as the hospital resources necessary to treat the condition.

In 2012 a tailored P4P program was introduced to control the payment amount provided to each hospital on the basis of their effectiveness. Specifically, on top of their annual budget each hospital receive a financial incentives based on a weighted mean of the hospital's evaluated outcomes. According to this measure the best-performing hospital receives an increment of 2% of its annual budget, the least-performing one gets a penalty of 2%, whereas all the others receive an amount between the interval $[-2\%, +2\%]$ and proportional to the distance between their score and the score of the least-performing hospital (see page 84 of [3], and [4]).

The evaluation of the hospitals outcome measures is assessed on 9 wards exogenously selected by the regional health-care management, namely cardiology, cardiosurgery, neurosurgery, neurology, oncology, general medicine, urology, or-thopaedic, and surgery. These wards have been chosen according to their coverage within the hospitals, the inclusion of both medical and surgical disciplines as well as the level of specialization, i.e. cardiosurgery and neurosurgery, and constitute the treatment group in the P4P evaluation analysis, whereas the other hospital wards not involved in the program belong to the control group. Further details on the policy introduction can be found in the regional resolution [3]. It is interesting to note that the evaluation is based on the selected wards, whereas the incentive is provided to the hospital as a whole, as typical of P4P programmes in healthcare [22]. Each hospitals have then a large accountability on how to allocate the incentive payments. Typically, provider institutions allocate the financial resources to make general improvements in the service delivered, and in particular related to the performance measures. In the case of the Lombardy region, it is also possible that the physicians and/or nurses working in the treated wards received a direct bonus to drive

performance improvement. This is however bound to vary across hospitals, so we do not expect to see the impact of this in our policy evaluation.

As in the evaluation of any policy, a choice needs to be made about which health outcome to use for quantifying the impact of the P4P program. In many studies, a single outcome is considered, such as overall mortality in England [98]. In addition, the evaluation of P4P programmes is often confined to specific clinical conditions, such as acute myocardial infarction (AMI), coronary artery bypass graft surgery (CABG), heart failure, pneumonia, and hip/knee replacement [42, 56, 61, 90, 98]. In contrast to these studies, we evaluate the performance of the hospitals by considering five outcome measures, namely overall mortality (in-hospital mortality + 30 days after discharge), number of transfers to a different hospital, number of discharges against medical advice, number of returns to the surgery room, and number of repeated hospitalisations or readmissions. The choice of these outcomes was based both on their popularity in the scientific literature, i.e. mortality and readmissions, and on the necessity of driving hospitals towards a reduction in the number of adverse outcomes, such as voluntary discharges, return to the surgery room and transfers to a different hospital. These outcomes measure have been previously identified by the Lombardy regional healthcare directorate to systematically evaluate the performance of the hospitals in terms of the quality supplied. More details of this process are given in [16] and in the regional resolution (page 4 of [4]).

2.1.3 Data

The database was gathered from the Lombardy healthcare information system. Data were collected on patients admitted to 142 hospitals during the four years 2010–2013 (two before and two in the policy-on period). In this period the hospitals provided 3,581,389 hospitalisations, coded in the available hospital discharge chart. In our analysis, we included patients admitted for acute care and we excluded patients living outside the region, patients younger than two years old or patients hospitalized in day-hospital, rehabilitation or palliative treatments. We used variables both at the patient and ward/hospital level. At the patient level, there is information on their gender, age, number of transit to the intensive care unit during hospitalization, the weight of the financial reimbursement corresponding to the patient's disease, and the comorbidity index. The latter is measured as in [38] and indicates the presence of one or more additional diseases or disorders co-occurring with a primary disease or disorder. At the hospital level, we know whether the hospital is affiliated to a medical school in which medical students receive practical training, whether the hospital is mono-specialistic or general, and whether there is presence of high-technology instrumentation in the ward. Finally, we include the hospitals' ownership, which categorizes the hospital as private for profit, private not-for-profit or public, and we distinguish wards whose prevalent activity is surgical from the medical ones. The effectiveness of the policy is evaluated over the five health outcomes described in the previous section, namely mortality, readmissions, transfers, returns, and voluntary discharges. We should clarify that the outcome return to the surgery room can be evaluated only for the surgical wards.

Table 2.1 reports the average (and the standard deviations in brackets) of the variables in the dataset by treatment and across the four years of the study (two pre and two post policy). It appears that the mix of patients within the treated and untreated wards is relatively constant over time, but that there are differences between the two groups. In particular, patients that are admitted to the treated wards are on average older than those admitted to the untreated ward. In addition, the treated wards consider higher risk patients than the untreated wards in terms of DRGs weight, number of comorbidities and intensive care treatment. The percentage of comorbidities (roughly 30%) is however still relatively small compared to other countries e.g. 0.69% in Northern Ireland in 2011/2012 [80]. This is justified by the coding rules that affect the healthcare system in Lombardy, whereby only the comorbidities directly connected with the treated DRGs are registered. Considering the variables related to the hospitals and the wards, we observe that the overall composition of the hospitals has not changed during the policy period, with surgical wards covering around 51% of the overall admissions. Moreover, 71% of the hospitalizations are provided by the public hospitals, whereas 30% of the patients are admitted to a

TABLE 2.1: Sample means and standard deviations in brackets for the covariates in the study from the Lombardy hospital inpatient stays for each year before and after the policy introduction.

	Untreated				Treated			
	Pre-policy		Post-policy		Pre-policy		Post-policy	
	2010	2011	2012	2013	2010	2011	2012	2013
Patient								
MALE	0.2589	0.2613	0.2646	0.2673	0.5399	0.5413	0.5397	0.5383
	(0.43)	(0.43)	(0.44)	(0.44)	(0.49)	(0.49)	(0.49)	(0.49)
AGE	46.076	46.585	46.973	47.212	64.526	64.877	65.054	65.384
	(21.1)	(21.1)	(21.2)	(21.3)	(18.7)	(18.5)	(18.6)	(18.5)
DRGWEIGHT	0.892	0.9127	0.9139	0.919	1.2974	1.3252	1.3167	1.3277
	(0.81)	(0.84)	(0.83)	(0.85)	(1.12)	(1.15)	(1.12)	(1.13)
COMORBIDITY	0.2379	0.2128	0.2156	0.2099	0.4082	0.3303	0.325	0.3121
	(0.58)	(0.55)	(0.56)	(0.55)	(0.72)	(0.66)	(0.65)	(0.64)
INTCARE	0.015	0.0164	0.017	0.0174	0.0644	0.0676	0.0677	0.0687
	(0.12)	(0.12)	(0.12)	(0.13)	(0.24)	(0.25)	(0.25)	(0.25)
Ward/Hospital								
TECHNOLOGY	0.8585	0.8588	0.8614	0.8683	0.8079	0.807	0.8111	0.8119
	(0.34)	(0.34)	(0.34)	(0.33)	(0.39)	(0.39)	(0.39)	(0.39)
TEACHING	0.2684	0.2708	0.2754	0.2734	0.2455	0.2456	0.2471	0.2456
	(0.44)	(0.44)	(0.44)	(0.44)	(0.43)	(0.43)	(0.43)	(0.43)
SPECIALISED	0.052	0.0474	0.0482	0.049	0.0387	0.0386	0.0406	0.0393
	(0.22)	(0.21)	(0.21)	(0.21)	(0.19)	(0.19)	(0.19)	(0.19)
SURGICAL	0.5637	0.5535	0.5646	0.562	0.5088	0.4884	0.4942	0.487
	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.5))	(0.49)
OWN:NOPROFIT	0.0758	0.0765	0.077	0.0793	0.0947	0.0948	0.0975	0.096
	(0.26)	(0.26)	(0.26)	(0.27)	(0.29)	(0.29)	(0.29)	(0.29)
OWN:PROFIT	0.1376	0.1373	0.1346	0.1264	0.2314	0.2354	0.2308	0.2327
	(0.34)	(0.34)	(0.34)	(0.33)	(0.42)	(0.42)	(0.42)	(0.42)
OWN:PUBLIC	0.7866	0.7862	0.7884	0.7943	0.6739	0.6698	0.6717	0.6713
	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.49)	(0.5))	(0.49)
Outcomes								
TRANSFERS	0.0056	0.0052	0.0036	0.0035	0.0127	0.0127	0.0053	0.0051
	(0.07)	(0.07)	(0.06)	(0.05)	(0.11)	(0.11)	(0.07)	(0.07)
RETURN	0.0592	0.0632	0.0099	0.0108	0.0431	0.0443	0.0154	0.0161
	(0.23)	(0.24)	(0.09)	(0.10)	(0.20)	(0.20)	(0.12)	(0.12)
MORTALITY	0.0268	0.0276	0.029	0.0273	0.0593	0.0608	0.0611	0.0601
	(0.16)	(0.16)	(0.16)	(0.16)	(0.23)	(0.23)	(0.23)	(0.23)
READMISSIONS	0.1216	0.1149	0.1117	0.1091	0.1335	0.1277	0.1211	0.1111
	(0.32)	(0.31)	(0.31)	(0.31)	(0.34)	(0.33)	(0.32)	(0.31)
VOLDISCH	0.0084	0.0085	0.0082	0.0084	0.0088	0.0081	0.0076	0.007
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.08)

private provider (20% in the for profit hospitals and 9% in the not-for-profit). With regards to the health outcome measures, three out of the five outcomes, namely transfers, return to the surgery room and readmissions, show a reduction after the introduction of the P4P program.

2.2 Methods

2.2.1 The DID approach

A before/after analysis, e.g. Hospital Quality Incentive Demonstration in USA [42], or a comparison analysis between participants and non-participants e.g. Quality Bonus System programme in Estonia [104] or Payment for Public Health Objectives programme in France [28], are common approaches for policy evaluation. When before/after analysis is considered, the average outcome is compared before and after the treatment in the treatment group only. The simplicity of

this approach comes at the expenses of the validity of the design study, as a time trend in the outcome may confound the effect of the treatment. One can compare the average difference in the outcome measure post treatment, between the treatment and control group, ignoring what happened in the pre-treatment period. In this case, the true treatment effect can be confounded by permanent differences in the treatment and control group that existed prior to any treatment. For these reasons, the DID method represent a stronger approach in terms of policy evaluations, and as such should be the preferred method for evaluating changes in health care policies [32].

The DID estimator is defined as a difference between the difference in the average outcome in the treatment group before and after the treatment and the same difference taken in the control group. In its simplest form, this can be achieved via a regression model which considers the treatment and the post policy factors only. For a continuous response, this is given by

$$Y = \theta_0 + \theta_1 \text{POST} + \theta_2 \text{TREATMENT} + \theta_3 \text{POST} \cdot \text{TREATMENT} + \epsilon.$$

The model has the following conditional expectations:

$$\begin{aligned} &E(Y|X, \text{POST} = 0, \text{TREATMENT} = 0) = \theta_0, \\ &E(Y|X, \text{POST} = 1, \text{TREATMENT} = 0) = \theta_0 + \theta_1, \\ &E(Y|X, \text{POST} = 0, \text{TREATMENT} = 1) = \theta_0 + \theta_2, \\ &E(Y|X, \text{POST} = 1, \text{TREATMENT} = 1) = \theta_0 + \theta_1 + \theta_2 + \theta_3, \end{aligned} \tag{2.1}$$

from which

$$\begin{aligned} &\text{D}_1 = E(Y|X, \text{POST} = 1, \text{TREATMENT} = 1) - E(Y|X, \text{POST} = 0, \text{TREATMENT} = 1) = \theta_1 + \theta_3 \\ &\text{D}_2 = E(Y|X, \text{POST} = 1, \text{TREATMENT} = 0) - E(Y|X, \text{POST} = 0, \text{TREATMENT} = 0) = \theta_1 \\ &\text{DID} = \text{D}_1 - \text{D}_2 = \theta_3 \end{aligned} \tag{2.2}$$

provide the differences in the expected outcome before and after the policy introduction in the treatment group, control group and their difference, respectively. Therefore the changes in outcome which are related to the policy introduction beyond background trends can be estimated from the double difference between the treated an the control group.

As detailed in [1], the DID estimator requires that the trends in outcomes between the treated and comparison groups are the same prior to the intervention i.e. parallel trend assumption. If true, it is reasonable to assume that these parallel trends would continue for both groups even if the program was not implemented. Moreover, any events occurring during or after the time the policy changed are assumed to equally affect the treatment and comparison groups. Thus, ideally, the only difference between the comparison group and the treatment group would be the exposure to the policy.

2.2.2 The econometric model for the P4P policy evaluation

We test the effect of the policy using a DID approach on data between 2010 and 2013, i.e. two year pre- and two year post-policy. To justify the suitability of this approach, the following considerations are needed:

1. The wards are split into a *treatment* group, i.e. the 9 wards that are used for the hospital evaluation, and a *control* group, i.e. the remaining wards. The allocation of the wards in one of these groups was made exogenously prior to the introduction of the policy [3]. There is an underlying assumption here that, although the incentive is provided to the hospital as a whole, the incentive is dictated only by the performance of the wards *treated*. Combined with the fact that the individual wards operate autonomously, the *untreated* wards can be considered as an independent

group. A similar analysis was conducted by [98], where the treatment and control groups are defined within each hospital on the basis of selected diagnoses.

2. Units do not switch between the control and the treatment group: improvements in performance of the control group do not affect the financial incentives gained by the hospital. We will however test whether there is evidence of a distortion of the hospital behaviour aimed at inflating the performance evaluation, such as the lift of resources in favour of the treated wards.

3. Any macro changes affect both groups equally and differences between the treatment and the control group remain constant in the absence of treatment, i.e. parallel trend prior to treatment. The check of this assumption is going to be discussed later in the results section. Of notice is also the fact that the regional resolution was formally announced in December 2011 [3], and applied from early January 2012 [4]. Thus, hospitals had no possibility to anticipate changes.

As discussed in the previous section, the policy evaluation is based on five health outcomes. Given the mix of patients in the different wards, the outcomes are first adjusted by patients characteristics via the use of a multilevel logistic mixed effect model [43, 92]. This model allows to account for the hierarchical structure of the data whereby patients are clustered into wards and wards are nested into hospitals. In addition, the longitudinal structure of the data means that a time effect is also to be expected. In detail, let Y_{pwht_m} represent a binary health outcome for patient p (with $p = 1, \dots, P_{wht_m}$) in the ward w (with $w = 1, \dots, W_{ht_m}$), belonging to the hospital h (with $h = 1, \dots, H_t$), hospitalized at time t_m (month of the years $t = 2010, \dots, 2013$). Let π_{pwht_m} be the conditional probability of Y_{pwht_m} being equal to 1. We consider the logistic regression mixed model

$$\text{logit}(\pi_{pwht_m}(x,u)) = \log\left(\frac{\pi_{pwht_m}(x,u)}{1-\pi_{pwht_m}(x,u)}\right) = \theta x_{pwht_m} + u_{wht_m} + u_{ht_m}, \tag{2.3}$$

where $\theta = (\theta_0, \theta_1 \dots, \theta_P)^T$ is a vector of coefficients for the $x_{pwht_m} = (1, x_1, \dots, x_P)$ patient-level covariates, and u_{wht_m} is the random effects for the ward w nested within hospital h at time t_m capturing the latent heterogeneity of the wards, whereas u_{ht_m} is the random effects capturing the latent heterogeneity of the hospital h at time t_m. In particular, u_{wht_m} and u_{ht_m} are independent and identically distributed i.e. $N(0, \sigma^2_{u_{wht_m}})$ and $N(0, \sigma^2_{u_{ht_m}})$ respectively, and are assumed to be uncorrelated with the regressors.

The model in Equation 2.3 returns the patients' predicted probabilities

$$\hat{\pi}_{pwht_m}(x,u) = \frac{\exp(\theta x_{pwht_m} + u_{wht_m} + u_{ht_m})}{1 + \exp(\theta x_{pwht_m} + u_{wht_m} + u_{ht_m})}, \tag{2.4}$$

which we collapse at the ward level over time in order to obtain the average predicted health outcome

$$\text{HO}_{wht_m} = \frac{\sum_{p \in P_{wht_m}} \hat{\pi}_{pwht_m}(x,u)}{|P_{wht_m}|}, \tag{2.5}$$

where P_{wht_m} is the set of patients admitted in the ward w of the hospital h in the month m ($m = 1, \dots, 12$) of the year t and $|P_{wht_m}|$ is the cardinality of this set.

The aim is now to quantify the policy effect on the basis of the five (adjusted) health outcomes. As we anticipate a correlation between the five health outcomes, we consider a multivariate DID model, rather than a separate model for each outcome. In this way, we are able to quantify the overall effect of the policy across all health outcomes, as well as at the individual level. Let then $\text{HO}^{(\lambda)}_{wht_m}$ denote the health outcome λ, namely readmissions ($\lambda = 1$), mortality ($\lambda = 2$), return to the surgical room ($\lambda = 3$), transfers ($\lambda = 4$) and voluntary discharges ($\lambda = 5$), at month m of year t ($t = 2010, \dots, 2013$) of

ward w ($w = 1, \dots, W_h$) belonging to hospital h (with $h = 1, \dots, H$). We consider the following multivariate mixed model:

$$\begin{aligned} \text{HO}^{(\lambda)}_{wht_m} = u^{(\lambda)}_h + \theta^{(\lambda)}_1 \text{TREATED}_{wh} + \sum_{j=2011}^{2013} \theta^{(\lambda)}_{2j} I(j=t) + \\ \sum_{j=2011}^{2013} \theta^{(\lambda)}_{3j} \left(I(j=t) \cdot \text{TREATED}_{wh}\right) + \theta^{(\lambda)}_4 \text{MONTH}_{t_m} + \epsilon^{(\lambda)}_{wht_m}, \end{aligned} \tag{2.6}$$

where the dummy variable TREATED_{wh} indicates whether the ward w within the hospital h is in the treatment group or not, the indicator variable $I(j = t)$ indexes the four years of the study (two pre and two post policy), with 2010 set as reference category, MONTH is a continuous variable, taking values 1 to 48 and added to correct for a possible seasonality effect, $u^{(\lambda)}_h$ is the random hospital effect for outcome λ, and the error $\epsilon^{(\lambda)}_{wht_m} = (\epsilon^{(1)}_{wht_m}, \dots, \epsilon^{(5)}_{wht_m})$ has a multivariate distribution $\epsilon_{wht_m} \sim \text{N}(0, \Sigma)$, with the covariance Σ accounting for possible dependencies between the different outcomes. The parameter $\theta^{(\lambda)}_{3j}$ is of interest in this model. Under the assumption of a parallel trend pre-policy, we expect $\theta^{(\lambda)}_{3,2011} = 0$ for all outcomes, whereas the parameters $\theta^{(\lambda)}_{3,2012}$ and $\theta^{(\lambda)}_{3,2013}$ represent the DID of average outcomes between the treated and control wards from the pre to the post-policy years. The two different parameters for the post-policy period let us detect whether the impact of the policy was immediate in the first year of its introduction or whether it was delayed in the second year [11]. This model allows us to detect the effect of the policy across all wards.

A second objective of the study is to detect whether the reaction to the P4P adoption is different depending on the ward's type. In particular, we group all wards into two types: surgical and medical, and extend the model in Equation 2.6 to:

$$\begin{aligned} \text{HO}^{(\lambda)}_{wht_m} = u^{(\lambda)}_h + \theta^{(\lambda)}_1 \text{TREATED}_{wh} + \sum_{j=2011}^{2013} \theta^{(\lambda)}_{2j} I(j=t) + \\ \sum_{k=1}^{2} \theta^{(\lambda)}_{3k} I(k = \text{SURGICAL}_{wh}) + \sum_{j=2011}^{2013} \left(\theta^{(\lambda)}_{4j} I(j=t) \cdot \text{TREATED}_{wh} \right) + \\ \sum_{j=2011}^{2013} \sum_{k=1}^{2} \left(\theta^{(\lambda)}_{5jk} I(j=t) \cdot I(k = \text{SURGICAL}_{wh}) \right) + \sum_{k=1}^{2} \left(\theta^{(\lambda)}_{6k} I(k = \text{SURGICAL}_{wh}) \cdot \text{TREATED}_{wh} \right) + \\ \sum_{j=2011}^{2013} \sum_{k=1}^{2} \left(\theta^{(\lambda)}_{7jk} I(j=t) \cdot I(k = \text{SURGICAL}_{wh}) \cdot \text{TREATED}_{wh} \right) + \theta^{(\lambda)}_8 \text{MONTH}_{t_m} + \epsilon^{(\lambda)}_{wht_m}, \end{aligned} \tag{2.7}$$

with the variable SURGICAL defined as 1 if the prevalent activity of the ward is surgical and 0 otherwise. In this model, the DID parameters $\theta^{(\lambda)}_{7,2012,k}$ and $\theta^{(\lambda)}_{7,2013,k}$ are of interest as they represent the differences in average outcomes between the surgical treated wards and the surgical control wards, from the pre to the post policy period and with respect to the medical wards which are taken as the reference category. For this model, we do not consider the health outcome returns to the surgery room as this is observed only for the surgical wards.

Finally, in the results section, we also consider a similar model for the detection of possible different reactions to the P4P adoption depending on the type of hospital ownership. In particular, we compare private for-profit, private not-for-profit and public hospitals. Due to the more strict budget constrains for private hospitals, these hospitals may react more actively to the policy than public ones. Furthermore, private for-profit hospitals are more oriented towards profit than the other hospitals and may therefore be more driven to increase their outcome measures in order to obtain a financial reward.

2.3 Policy evaluation

In this section, we use the models described above to evaluate the impact of the introduction of the P4P policy in Lombardy. Table 2.2 reports the fixed effects estimates of the model in Equation 2.6. As all outcomes are constrained to be between 0 and 1, the parameter estimates and the p-values are computed by a non-parametric bootstrap approach. For this, we use a method specifically developed for multilevel modelling [19, 109].

TABLE 2.2: Parameters estimates for the fixed part of the multivariate mixed model in Equation 2.6.

	MORTALITY	READMISSIONS	RETURN	TRANSFERS	VOL. DISCH.
MONTHS	0.001	-0.001	0.001	-0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
TREATED	0.02***	0.004***	-0.037***	0.006***	0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
$YEAR_{2010}$	0.044***	0.13***	0.084***	0.009***	0.009***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
$YEAR_{2011}$	0.044***	0.125***	0.082***	0.008***	0.008***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
$YEAR_{2012}$	0.045***	0.122***	0.021***	0.006*	0.008**
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)
$YEAR_{2013}$	0.041***	0.118***	0.022***	0.005	0.008**
	(0.004)	(0.004)	(0.006)	(0.004)	(0.004)
TREATED·$YEAR_{2011}$	0.002	0.001	0.002	0.001	-0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
TREATED·$YEAR_{2012}$	0.001	-0.005***	0.026***	-0.005***	-0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)
TREATED·$YEAR_{2013}$	0.005***	-0.011***	0.025***	-0.005***	-0.001
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2.3.1 Testing the assumptions of a DID approach for policy evaluation

Table 2.2 shows how the parameters $\theta^{(\lambda)}_{3,2011}$ of the interaction between TREATED and $YEAR_{2011}$ are not significantly different from zero. This provides evidence in favour of the parallel trend assumption for each individual health outcome, i.e. the differences between the average outcome of the treatment and control group are constant prior to the introduction of the policy. This assumption is needed in order to evaluate the impact of the policy using a DID approach. As we require a parallel trend to be satisfied for all health outcomes simultaneously, we use a multivariate analysis of variance test (MANOVA) to test the null hypothesis $H_0 : \theta^{(1)}_{3,2011} = \dots \theta^{(5)}_{3,2011} = 0$ under the model in Equation 2.6. The Wilks' lambda statistics returns a p-value of 0.2676, which provides further evidence in support of the parallel trend assumption across all health outcomes.

Given that the incentive is provided to the hospital as a whole, it is also necessary to test whether the introduction of the P4P may have had a negative spillover effect between the treated and the untreated wards. This would violate the assumption of independence between the two groups and thus bias the policy evaluation. Although within each ward the physicians and nurses detain managerial freedom on whether and how to treat the patients, spillover effects could take the form of hospitals lifting resources in favour of the treated wards to the expense of the untreated wards. To this aim, we assess whether there has been a difference in the total number of hours worked by physicians and nurses within each hospital between the treated and the untreated wards from the year 2011 (pre-policy) to 2012 (post-policy). We consider 58 hospitals which have a balanced proportion of treated/untreated wards.

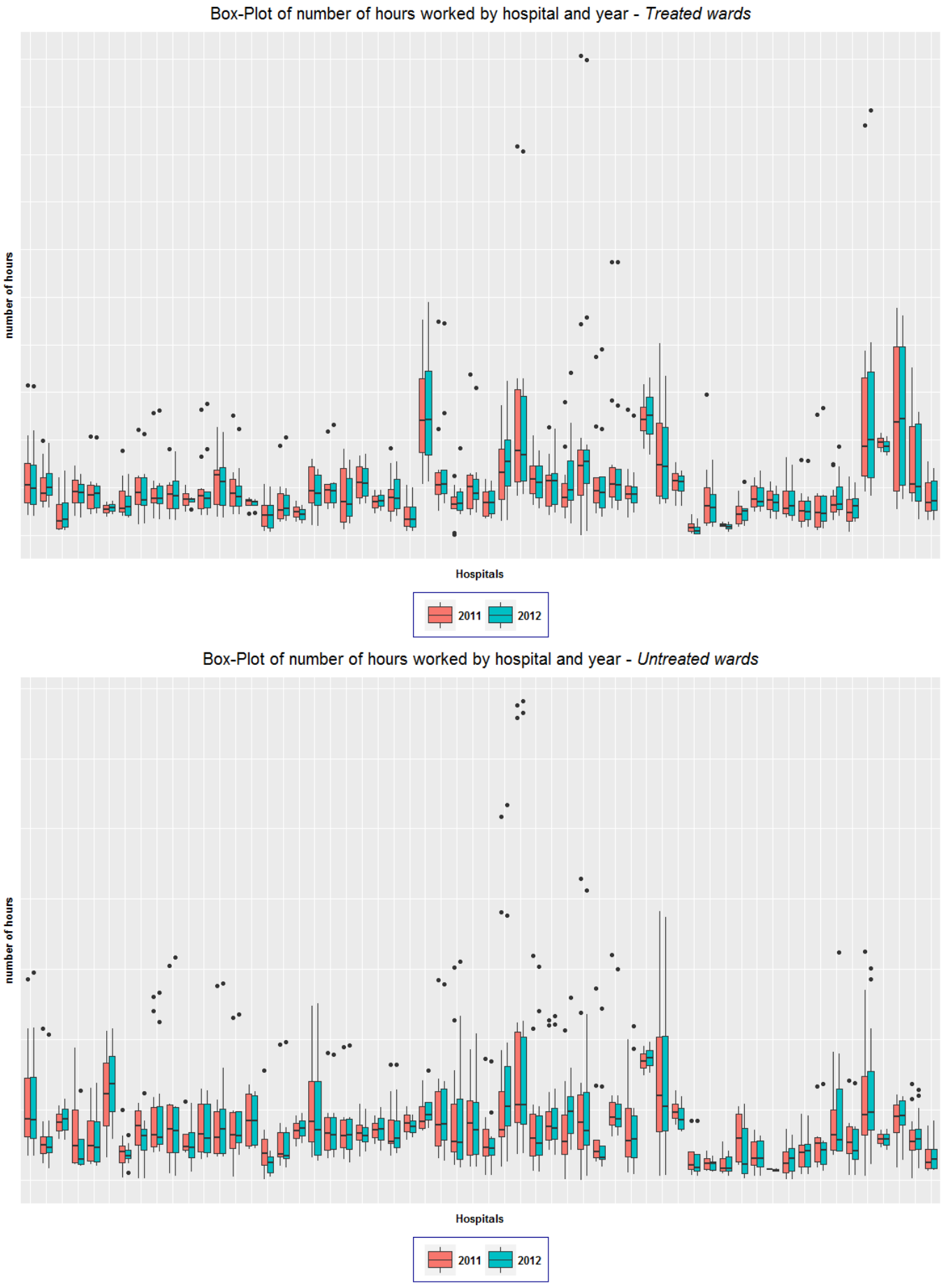


FIGURE 2.1: Box-plot of the number of hours worked by physicians and nurses across hospitals before and after the policy introduction for the treated (top) and untreated (bottom) wards.

Figure 2.1 shows the box-plot of the number of hours worked by hospital and year. The figure shows how, within each hospital, the number of hours worked is stable across the two groups and between the pre and post-policy period, suggesting that no shift of resources occurred, at least at the level of labour. This is supported by a non-significant p-value for the year-treatment interaction term (p-value=0.812) from a Negative Binomial generalised linear model (GLM) which also includes hospitals fixed effects. In addition to the allocation of resources, another possible spillover effect could result from the sharing of technological resources between the different wards. This may have an impact on surgical outcomes, such as the return to the surgery room in our case. We have no data to evaluate this, but we will take this into consideration when interpreting the results of the policy evaluation analysis.

Together with the spillover effects mentioned above between wards within the same hospital, the different providers may have also reacted to the policy by avoiding to treat high risk patients [60]. In order to check for this potential distortion, we have analysed whether the cream skimming index, calculated as in [14], changed significantly between the pre and the post policy period. As above, we restrict the analysis to the hospitals which have a balanced proportion of treated/untreated wards and we perform the pre-post analysis separately for the treated and untreated groups. Using a multiple regression model, we find only four hospitals (out of 58) with a significant negative interaction with the post-policy term, two for the treated wards (p-values=$4.54E-08$, and 0.0025) and two for the untreated ones (p-values= 0.02, and 0.0314). Thus, we conclude that overall the hospitals show no evidence of a gaming behaviour in selecting the mix of patients in the post-policy period.

2.3.2 Do the hospitals react positively to the policy?

We are now in a position to evaluate the impact of the P4P policy by considering the estimates of the coefficients of the interaction between the treatment variable and the post-policy years in Table 2.2, i.e. $\theta^{(\lambda)}_{3,2012}$ and $\theta^{(\lambda)}_{3,2013}$. As all health outcomes are improved if they are reduced, a significant and negative coefficient for these interactions would mean that the P4P introduction had a positive effect on quality. This result is confirmed for readmissions ($\theta_{3,2012}$=-0.0051, $\theta_{3,2013}$=-0.0112) and transfers ($\theta_{3,2012}$=-0.0046, $\theta_{3,2013}$=-0.0047). This is a clear signal that the hospital activity was modified as a result of the P4P introduction, as both readmissions and transfers are directly affected by the hospital organization. In particular, the results show that the P4P program may have reduced the hospital attitude of readmitting patients in order to increase the number of the DRGs provided [14]. The reduction in the transfers of the patients between hospitals in the treated wards is also particularly encouraging, considering that transfers are directly linked to the patient safety and continuity of care.

In order to further quantify the impact of the policy and to confirm the significance of the results on the health outcomes in absolute terms, Figure 2.2 plots the marginal effects of each health outcome in Equation 2.6 for treated and untreated wards and over the observation period [5, 57]. As well as verifying the parallel trend in the pre-policy period, the plots show a clear improvement for readmissions and transfers. In particular, there is an absolute difference of 0.91% and 1.52% in the average number of readmissions between the treated and untreated wards in the year 2012 and 2013, respectively, and of 0.31% in the year 2011, whereas there is a difference of 0.19% and 0.18% in the average number of transfers between the treated and untreated wards in the year 2012 and 2013, respectively, and of 0.72% in the year 2011. This leads to DID reductions of 0.60% (readmissions) and 0.53% (transfers) in 2012 compared to 2011 and a further reduction of 0.61% (readmissions) and 0.01% (transfers) in 2013. The predicted percentages of reduction correspond to a P4P-related saving of 4,324 readmissions and 4,295 transfers in the treated wards in 2012 and a further reduction of 4,871 readmissions and 157 transfers in 2013.

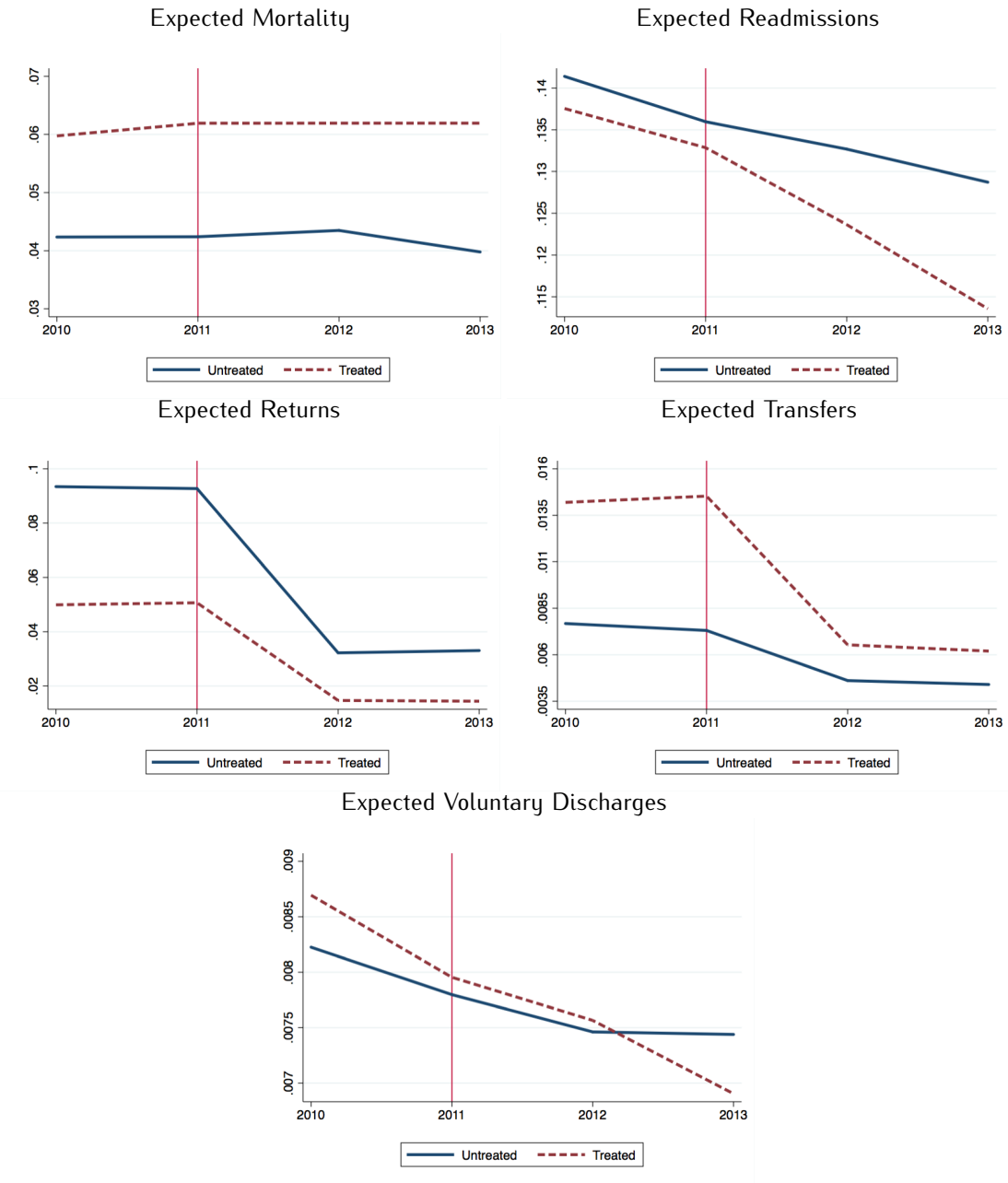


FIGURE 2.2: Marginal effects of all health outcomes per year and treatment for the model in Equation 2.6.

The picture for the other three health outcomes is more complex than for transfers and readmissions. The average number of returns to the surgery room seems to increase in the treated wards more than in the untreated after the introduction of the policy, as $\theta_{3,2012}$ and $\theta_{3,2012}$ are positive and significant. This is shown in Figure 2.2, which, on the other hand, shows also how the P4P incentives improve the performance for both the treated and untreated wards. This is an interesting result, suggesting that the managerial impact in the hospital organization caused by the adoption of the P4P program has changed the overall hospital performance with regards to the surgical activity. A possible explanation to this could be given by a spillover effect between the treated and the untreated wards, as all wards may benefit from potentially improved technology in the surgery room.

For the other two health outcomes, voluntary discharges and mortality, the DID coefficients of $\theta_{3,2012}$ and $\theta_{3,2013}$ are not significantly different from zero. Figure 2.2 shows how the number of voluntary discharges decreases already before the P4P introduction. With regards to mortality, it is reasonable to believe that, when hospitals are checked for effectiveness

on more than one output, they will focus on those outcomes that are easily measurable. This is observed by [78] in the context of a competition analysis. From this point of view, readmissions, transfers and return to the surgery room represent well-measured outcomes. Hence it is possible that hospitals have focussed their efforts on those easily measured and better observable activities in order to increase their performance and then gain financial rewards.

2.3.3 Do surgical and medical wards react differently to the policy?

We investigate the policy effect with regards to the different wards, by evaluating whether surgical and medical wards reacted differently to the policy. We fit the model in Equation 2.7 to the data in order to answer this question. The results, omitted in full for brevity, show evidence of a differential impact of the P4P introduction for the two health outcomes that were significant in the global analysis above. In particular, there is evidence that the P4P program impacted more on the medical wards than on the surgical ones in terms of number of readmissions ($\theta_{7,2012}$=0.008, p-value=0.0102; $\theta_{7,2013}$=0.0307, p-value=<.0001) and number of transfers ($\theta_{7,2012}$=0.0117, p-value=0.0002, $\theta_{7,2013}$=0.012, p-value=0.0001). This is shown visually also by the marginal effects in Figure 2.3. This finding can be explained by the fact that the surgical healthcare pathways are more rigorous and more linked to fixed guidelines than those on medical hospitalizations, which instead tend to be more flexible and more dependent on managerial actions and hospital organization.

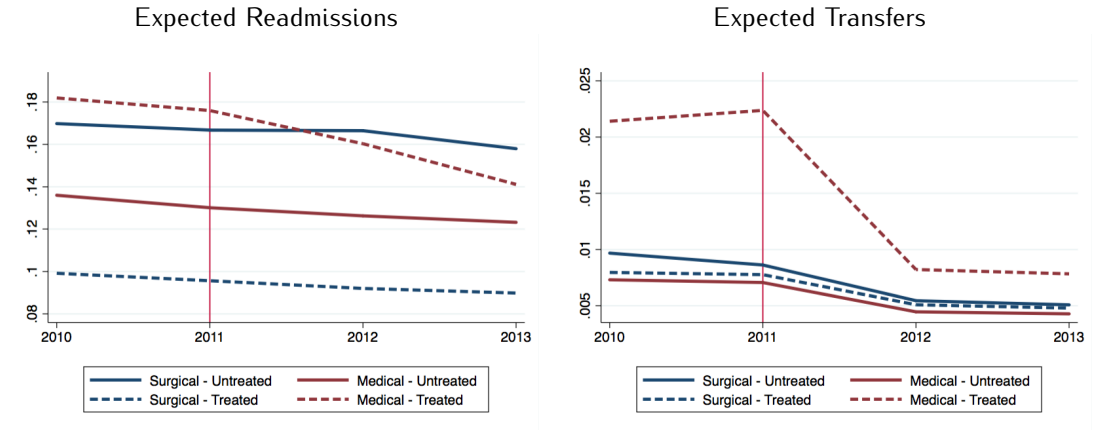


FIGURE 2.3: Marginal effects of readmissions and transfers per type of ward, year and treatment for the model in Equation 2.7.

2.3.4 Do private and public hospitals react differently to the policy?

Previous studies have found no dependency between hospital ownership and efficiency [12] or hospital ownership and competition [15], suggesting that the long term adoption of a quasi-market system in Lombardy has reduced the expected differences between the hospital types. For the first time in a P4P study, here we investigate the policy effect with regards to hospital ownership, by evaluating possible different reactions to the P4P program among the private (for-profit and not-for-profit) and public providers. To answer this question, we use a model like Equation 2.7, but with SURGICAL replaced by a variable representing the ownership type, where the public hospitals are taken as the reference category. Once again, the interactions $\theta^{(\lambda)}_{7jk}$ are of interest in this model. In line with the existing literature, the results show only limited evidence in support to a hypothesis of a different reaction: apart from readmissions in 2012 ($\theta_{7,2012,\text{not-for-profit}}$=-0.01964, p-value=0.0004; $\theta_{7,2012,\text{private}}$=-0.0096, p-value=0.0062), the interaction for readmissions in 2013 and all interactions for transfers, for both the private for profit and not-for-profit categories, are not statistically significant. This is an interesting result meaning that the monetary incentive is a valuable motivation to improve the quality of care of hospitals with all types of ownership and not only for the profit-maximizer providers, i.e. profit hospitals.

2.4 Conclusions

The P4P approach has been adopted in many countries in order to encourage improvements in the quality of healthcare by supplying financial incentives to healthcare providers. In this study, we evaluate the impact of a specific P4P program adopted in the Lombardy region (Italy) in 2012. Differently to previous studies, we perform the analysis considering the whole healthcare system, evaluating multiple health outcomes over a number of clinical areas. We analyse data over four years, two before (2010/2011) and two after (2012/2013) the implementation of the program. The policy was applied to all hospitals in the Lombardy region, but the incentive was calculated only on the basis of the performance of 9 wards. The fact that the selection of these wards was made exogenously, combined with the fact that we observe a parallel trend pre-introduction of the policy and that we have found no evidence of spillover effects between the treated and untreated wards in terms of allocation of resources, have led us to use a multivariate DID approach for the evaluation of the impact of the policy.

Our study shows that two out of the five health outcomes considered i.e. readmissions and transfers, support the hypothesis that the P4P introduction had a positive effect on quality. The picture for the other three health outcomes is more complex than for transfers and readmissions. Considering the returns to the surgery room, our results show that the P4P incentives improve the performance for both the treated and untreated wards. We speculate that this may be the result of improved technology in the surgery room which all the wards have benefit from. The last two health outcomes, voluntary discharges and mortality, did not show changes that can be attributed to the P4P adoption. This can be explained by considering the fact that when hospitals are checked for effectiveness on more than one output, they will focus on those outcomes which are more easily driven by a managerial intervention in order to improve their performance and to obtain the financial incentives. Moreover, our study shows that the medical wards have reacted to the P4P program more strongly than the surgical wards, whereas only limited evidence is found to suggest that the policy reaction was different across different types of hospital ownership. As anticipated by [23], overall the results show that the healthcare system in Lombardy was positively impacted by the P4P implementation: there is evidence of a reduction in some adverse health outcomes and of a general change in the hospital organization in order to improve the healthcare services provided to the citizens. Lastly, the evaluation study found no evidence of a distortion of the hospital behaviour aimed at inflating the performance evaluation, such as cream skimming behaviour.

This study has some implications. Firstly, Lombardy should extend the adoption of the P4P program across the whole regional healthcare system in order to improve the overall hospital activity. Secondly, given the positive impact of the P4P program in Lombardy, the adoption of a similar strategy is suggested to the other regional healthcare systems in Italy. This would stimulate improvements in quality for the regions that already perform relatively well, but, in particular, this would be an important incentive for these regions with a lower qualified healthcare system. The same could also apply to other countries.

Future work on the evaluation of P4P programs could explore additional aspects. Firstly, we could test the effect of the adoption of the P4P program by using more flexible models, such as via autoregressive time components. Secondly, it would be interesting to test the impact of the P4P program in terms of the number of intra-hospital infections and complications, or other outcomes directly related to the performance of the hospitals' physicians and the improvement of technology. Thirdly, it would be useful to conduct a comparative analysis between the Lombardy region and neighbouring regions which are not subjected to P4P programmes. This would help also in controlling for spillover effects between the treated and the untreated wards within the same hospital, such as those resulting from the sharing of common technology and resources. Fourthly, our analysis has focussed solely on the impact of the P4P programs on the hospital effectiveness. It would be interesting to extend the current analysis to understand whether the monetary incentive had an impact also on the hospital efficiency. Finally, we believe that further research is needed to assess the impact of P4P programs over a long time frame, as encouraged by [110].

Chapter 3

Linear models for counts via a Discrete Weibull distribution

Motivated by the lack of a unique, efficient, and flexible regression framework for the different types of count response, i.e. over- or under-dispersed, and excessive zeros, we develop regression models via a Discrete Weibull distribution. The analyses presented in this chapter and the one presented in chapter 4 have been conducted in R software [79].

3.1 The Discrete Weibull distribution and its properties

The Discrete Weibull distribution was introduced by [71], as a discrete form of a continuous Weibull distribution, similarly to the Geometric distribution, which is the discrete form of the Exponential distribution, and the Negative Binomial, which is the discrete alternative of a Gamma distribution. In some studies this is referred to as type I Discrete Weibull, as two other distributions were subsequently defined. The three distributions have been reviewed by [17] which point out the advantages of using the type I distribution, i.e. it has an unbounded support, differently to the type II distribution, and it has a more straightforward interpretation differently to the type III distribution. The probability mass function of a type I Discrete Weibull

$$Y|X \sim \text{Discrete Weibull}(q(x), \beta)$$

can be written as

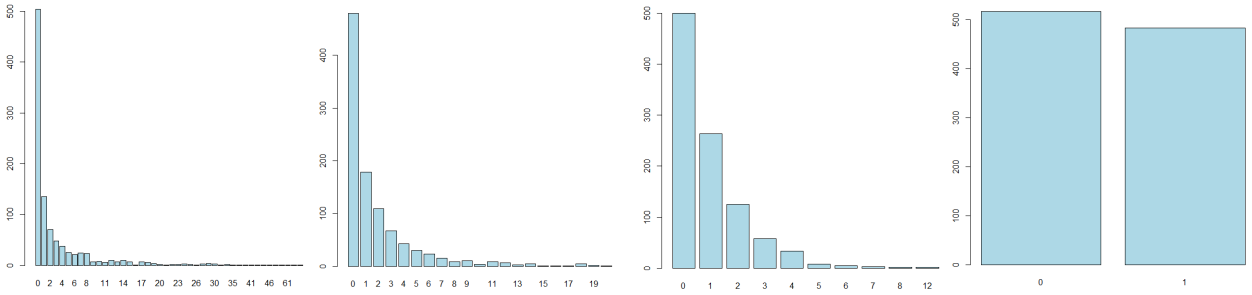
$$f(y; q(x), \beta) = q(x)^{y^\beta} - q(x)^{(y+1)^\beta}, \tag{3.1}$$

for $y = 0, 1, 2, \ldots$, with the real parameter $0 < q(x) < 1$, and the shape parameter $\beta > 0$. Thus, the cumulative distribution function can be written as

$$F(y; q(x), \beta) = \begin{cases} 1 - q(x)^{(y+1)^\beta} & y = 0, 1, 2, \ldots \\ 0 & y < 0. \end{cases} \tag{3.2}$$

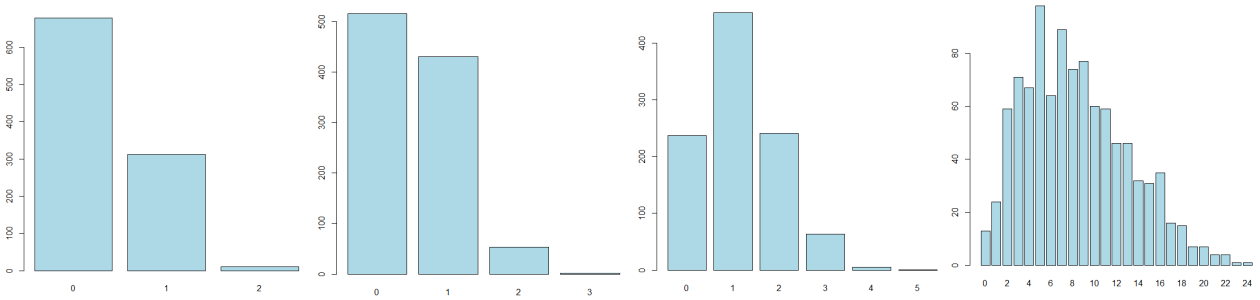
Figure 3.1 and Figure 3.2 shows how both the $q(x)$ and β parameter affect the shape of the Discrete Weibull distribution. Specifically, a value of β close to 0 leads to a highly skewed distribution, while a value of β close to ∞ reduces the

range of the count values of the response. Moreover, the $q(x)$ parameter quantify the probability of the response variable being a non-null value, i.e. $\Pr(Y=0|X)=1-q(x)$, thus $q(x)=1-\Pr(Y=0|X)=\Pr(Y>0|X)$.



DW q(x)=0.5, β=0.5 DW q(x)=0.5, β=0.7 DW q(x)=0.5, β=1 DW q(x)=0.5, β=4

FIGURE 3.1: Plot of the Discrete Weibull distribution for different values of β, and q(x)=0.5.



DW q(x)=0.3, β=2 DW q(x)=0.5, β=2 DW q(x)=0.75, β=2 DW q(x)=0.99, β=2

FIGURE 3.2: Plot of the Discrete Weibull distribution for different values of q(x), and β=2.

The mean of a Discrete Weibull with parameters q(x) and β as described in Equation 3.1 can be written as

$$\mathrm{E}(Y|X)=\sum_{y=1}^{\infty}(q(x))^{y^{\beta}}, \tag{3.3}$$

while the variance of a Discrete Weibull$(q(x),\beta)$ can be written as

$$\begin{aligned}\mathrm{var}(Y|X)&=\sum_{y=1}^{\infty}(2y-1)(q(x))^{y^{\beta}}\\&=2\sum_{y=1}^{\infty}y(q(x))^{y^{\beta}}-\mathrm{E}(Y|X).\end{aligned} \tag{3.4}$$

The distribution is connected to other well known distributions. In particular,

- The discrete Rayleigh distribution in [86] is a special case of a Discrete Weibull with $\beta=2$ and $q(x)=x\theta$.
- The Geometric distribution is a special case of a Discrete Weibull, with $\beta=1$ and $q(x)=1-p(x)$. Moreover, for the Geometric distribution the variance is always greater than its mean. Therefore, a Discrete Weibull with $\beta=1$ is a case of over-dispersion relative to Poisson, regardless of the value of $q(x)$. In particular, when $\beta=1$ and $q(x)=e^{-\lambda(x)}$, the distribution is the Discrete Exponential distribution introduced by [87].
- β can be considered as controlling the range of values of the variable. As $\beta\to\infty$, the Discrete Weibull approaches a Bernoulli distribution with probability $q(x)$.

As for quantiles, the τ-quantile of a Discrete Weibull of parameters $q(x)$ and β is given by the smallest integer $\mu^{(\tau)}$ for which $\Pr(Y = y|X \leqslant \mu^{(\tau)}) = 1 - (q(x))^{(y+1)^{\beta}} \geqslant \tau$. Thus, the Discrete Weibull distribution presents a closed form for its τ-quantile function, which is given by

$$\mu^{(\tau)} = \left\lceil \left(\frac{\log(1-\tau)}{\log(q(x))} \right)^{\frac{1}{\beta}} - 1 \right\rceil, \tag{3.5}$$

where $\lceil \cdot \rceil$ is the ceiling function. As a special case, the median of the Discrete Weibull is given by

$$\mu^{(0.5)} = \left\lceil \left(-\frac{\ln(2)}{\ln(q(x))} \right)^{\frac{1}{\beta}} - 1 \right\rceil. \tag{3.6}$$

The formulation in Equation 3.5 can be extended to non integers by removing the ceiling function, though since the Discrete Weibull takes positive values only, this will be valid only for $\tau \geqslant 1 - q(x)$.

3.1.1 Accounting for different types of dispersion

Dispersion in count data is formally defined in relation to a specified model being fitted to the data [18, 50]. In particular,

$$\text{VR} = \frac{\text{observed variance}}{\text{theoretical variance}}. \tag{3.7}$$

So VR is the ratio between the observed variance from the data and the theoretical variance from the model. Then the data are said to be over-/equi-/under- dispersed relative to the fitted model if the observed variance is larger/equal/smaller than the theoretical variance specified by the model, respectively. It is common to refer to dispersion relative to Poisson. In that case, the variance of the model is estimated by the sample mean. Thus, over-/equi-/under- dispersion relative to Poisson refers to cases where the sample variance is larger/equal/smaller than the sample mean, respectively. Since the theoretical variance of a Negative Binomial is always greater than its mean, the Negative Binomial regression model is the natural choice for data that are over-dispersed relative to Poisson. However, crucially, the Negative Binomial distribution cannot handle under-dispersed data. In contrast to this, Figure 3.3 shows how a Discrete Weibull distribution can handle data that are both over- and under- dispersed relative to Poisson.

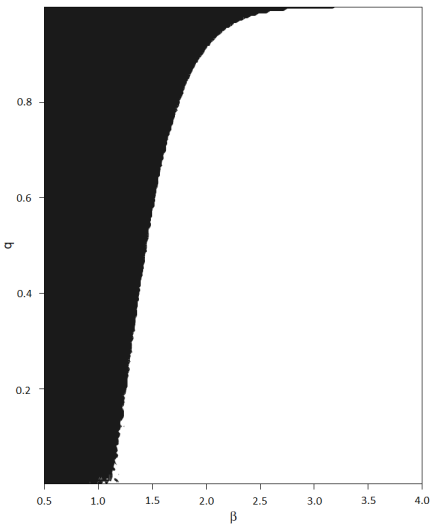


FIGURE 3.3: Ratio of observed and theoretical variance from a Poisson model, calculated from simulated Discrete Weibull models with parameters q(x) and β.

Specifically, the white area corresponds to values of dispersion less than 1, i.e. under-dispersed relative to Poisson, whereas the black area corresponds to over-dispersion. Moreover, the plot shows that:

- $0 < \beta \leq 1$ is a case of over-dispersion, regardless of the value of $q(x)$.
- $\beta \geq 3$ is a case of under-dispersion, regardless of the value of $q(x)$. In fact, the Discrete Weibull approaches the Bernoulli distribution with mean $p(x)$ and variance $p(x)(1-p(x))$ for $\beta \to \infty$.
- $1 < \beta < 3$ leads to both cases of over and under-dispersion depending on the value of $q(x)$.

3.2 The Discrete Weibull regression model

3.2.1 Linear regression model

There are a number of possible choices for linking the q and β parameters to linear predictors x. In particular,

- q depend on x via

$$\log\left(\frac{q(x)}{1-q(x)}\right) = x\theta,$$

  i.e. $\log(q(x)) = x\theta - \log(1+\exp(q(x)))$, where $x = (1, x_1, \ldots, x_P)$, $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$.

- q depend on x via

$$\log\left(-\log\left(q(x)\right)\right) = x\theta,$$

  where $x = (1, x_1, \ldots, x_P)$, $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$

- β depend on x via

$$\log(\beta(x)) = x\vartheta,$$

  where $x = (1, x_1, \ldots, x_P)$, $\vartheta = (\vartheta_0, \vartheta_1, \ldots, \vartheta_P)^T$, otherwise β is kept constant.

- both q and β depend on x.

While the parametrization which consider the logit link function has been exploited by [46], the analysis presented in this thesis is based on the $\log(-\log)$ link in $q(x)$. Thanks to this link we will show later in subsection 3.3.2 how this model formulation can be linked to that of a continuous Weibull regression models so that efficient implementations can be made available in `R` software. Moreover, in subsection 3.3.3 we will show how the analytical formula for the quantile facilitates the interpretation of the parameters of this model formulation. Regarding the β parameter, this chapter considers it fixed, that is we consider the model

$$\begin{aligned} \log\left(-\log\left(q(x)\right)\right) &= x\theta, \\ \log\left(\beta\right) &= \vartheta \end{aligned} \tag{3.8}$$

where $x = (1, x_1, \ldots, x_P)$, $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$, and ϑ takes real values. Later in chapter 3 we will also consider the dependence of β on x in order to capture more complex dependencies. Inference for these models is included in our `R` package `DWreg`.

3.2.2 Linear mixed regression model

It is possible to extend the linear formulation in Equation 3.8 with the inclusion of random effects. This leads to

$$\begin{aligned} \log(-\log(q(x,u)) &= x\theta + zu \\ \log(\beta) &= \vartheta \end{aligned} \tag{3.9}$$

where $x = (1, x_1, \ldots, x_P)$, $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^T$, $\vartheta = (\vartheta_0)$, and $z = (z_1, \ldots, z_Q)$ is the $(Q \times 1)$ design vector for the random effects $u = (u_1, \ldots, u_Q)^T$ which are assumed i.i.d. as Normal$(0, \sigma_u^2)$.

3.2.3 Excess zeros regression model

In addition to the cases of over- or under- dispersion, it is important to consider the presence of excessive zeros. In fact, the joint presence of excess zeros and long right tails are features common to many counts. Typically, an excess of zeros in the data reduces the mean of the response, thus inflating the dispersion index. Hence, it is important to consider a flexible distribution as the Discrete Weibull which not only can account for the excess of zero, but can also address potential over- or under- dispersion. Models such as zero-inflated or hurdle regression are employed when there is evidence of an excess of zeros in the data.

Zero inflated Discrete Weibull regression model As detailed in subsection 1.2.2, zero inflated models combines zeros coming from both a point mass at zero and a conditional count distribution. Thus, the zero inflated Discrete Weibull regression model with parameter q(x), β and $\pi(x)$ can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) + (1-\pi(x))(1-q(x)) & \text{for } y = 0 \\ (1-\pi(x))(q(x)^{y^\beta} - q(x)^{(y+1)^\beta}) & \text{for } y = 1, 2, 3, \ldots, \end{cases} \tag{3.10}$$

where $0 < \pi(x) < 1$ is the mixture parameter which is related to the set of covariates by

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x\gamma, \tag{3.11}$$

with $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_P)^T$.

Hurdle Discrete Weibull regression model Another possibility to model data with excessive zeros is hurdle regression, as detailed in subsection 1.2.2. The hurdle Discrete Weibull regression model with parameter q(x), β and $\pi(x)$ can be written as

$$Pr(Y|X) = \begin{cases} \pi(x) & \text{for } y = 0 \\ (1-\pi(x))\frac{(q(x)^{y^\beta} - q(x)^{(y+1)^\beta})}{q(x)} & \text{for } y = 1, 2, 3, \ldots. \end{cases} \tag{3.12}$$

Inference for these models is included in our `R` package `DWreg`.

3.3 Parameter estimation

In a regression framework, there are different methods to estimate the parameters, i.e. GLS, GEE or MCMC, as illustrated in [53]. Here, we use full maximum likelihood. This method is generally robust, and produces estimates that are asymptotically efficient and consistent.

3.3.1 Likelihood

Under a maximum likelihood approach, the parameters θ and ϑ of Equation 3.8, are estimated by directly maximising the likelihood function using any non-linear optimization tool. The likelihood function can be written as

$$L(y,x;\theta,\vartheta)=\prod_i f(y_i|x_i)=\prod_{i=1}^{n}\left((q(x_i))^{y_i^\beta}-(q(x_i))^{(y_i+1)^\beta}\right),$$

on the data $y=(y_1,\dots,y_n)$, and the maximum of which can be found numerically. This leads to the log-likelihood function

$$l(y,x;\theta,\vartheta)=\sum_{i=1}^{n}\log\left((q(x_i))^{y_i^\beta}-(q(x_i))^{(y_i+1)^\beta}\right).$$

The optimisation of this likelihood was originally implemented in the `R` package `DWreg` [108]. In the next section we discuss a faster alternative which also opens up the possibility for DW-inference for other regression models, such as the mixed model in Equation 3.9.

3.3.2 Link between discrete and continuous Weibull distribution

As introduced in section 3.1, the Discrete Weibull has been derived as the discrete analogues of a continuous Weibull distribution, i.e. see methodology-IV in [24]. In particular, the latter can be described as

$$Y|X\sim\text{Weibull}(\mu(x),\sigma),$$

with probability density function and cumulative density function defined by

$$\begin{aligned}f_{\text{W}}(y)&=f(y;\mu(x),\sigma)=\frac{\sigma}{\mu(x)}\left(\frac{y}{\mu(x)}\right)^{(\sigma-1)}\exp\left\{-\left(\frac{y}{\mu(x)}\right)^{\sigma}\right\}\qquad y\geq 0\\ F_{\text{W}}(y)&=F(y;\mu(x),\sigma)=1-\exp\left\{-\left(\frac{y}{\mu(x)}\right)^{\sigma}\right\}.\end{aligned}\tag{3.13}$$

Let us recall the probability mass function and the cumulative density function of a Discrete Weibull presented in Equation 3.1 and Equation 3.2, respectively, i.e.

$$\begin{aligned}f_{\text{DW}}(y)&=f(y;q(x),\beta)=(q(x))^{y^\beta}-(q(x))^{(y+1)^\beta}\qquad y=0,1,2,\dots\\ F_{\text{DW}}(y)&=F(y;q(x),\beta)=1-(q(x))^{(y+1)^\beta}.\end{aligned}$$

We consider the transformation from the continuous to the discrete case given by

$$\begin{aligned}\exp\left\{-\frac{1}{\mu(x)}\right\}&=q(x)\\ \sigma&=\beta,\end{aligned}$$

and by substituting this into

$$F_{\mathrm{W}}(y+1) - F_{\mathrm{W}}(y),$$

we obtain the $f_{\mathrm{DW}}(y)$ with parameters $\beta > 0$ and $0 < q(x) < 1$. Thus, the likelihood of a continuous Weibull distribution with interval censored data is equal to that of a Discrete Weibull distribution, i.e.

$$\prod_{i=1}^{n} \left(F_{\mathrm{W}}(y_i+1|x_i) - F_{\mathrm{W}}(y_i|x_i)\right) = \prod_{i=1}^{n} f_{\mathrm{DW}}(y_i|x_i).$$

From this,

$$\int_{y}^{y+1} f_{\mathrm{W}}(y)\,\mathrm{d}y = f_{\mathrm{DW}}(y)$$

shows that integrating between y and $y+1$ the probability density function of a continuous Weibull leads to the probability mass function of a Discrete Weibull distribution.

From these considerations, we can use available implementations for continuous Weibull regression models with interval censored data, such as the function `gamlss` in the `R` package `gamlss` [93], and the function `survreg` in the `R` package `survival` [100]. The `survreg` implementation is limited to simple regressions with the possibility of adding a simple random effects term, i.e. a frailty. In contrast to this, the `gamlss` implementation allows to include complex non-linear and multilevel models, so it will be chosen implementation for this thesis. The interval censored response variable can be created in `R` software by calling the `survival` package and make use of the `Surv` function with `type=interval2`.

Link between the parameter estimates and `gamlss` and `survreg` parametrisations We exploit the link between the Discrete Weibull with parameters $q(x)$ and β as presented in Equation 3.1, and the parametrisation of a continuous Weibull with parameters $\mu(x)$ and σ as presented in Equation 3.13 and implemented in `R` software within the `gamlss.dist` and `survival` package. In particular, the estimators of the parameters of the Discrete Weibull were derived by a direct transformation, whereas the standard errors were derived using a first-order Taylor expansion around the mean known as Delta method [75]. Table 3.1 shows these transformations, while the details on how to derive the standard errors of the parameters can be found in section A.1.

TABLE 3.1: Discrete Weibull model parameters and respective standard errors exploiting the `survreg` and `gamlss` parametrisation via a continuous Weibull distribution.

Estimates	Std. Errors
`survreg`	
$\hat{\beta} = \frac{1}{\hat{\sigma}}$	$\text{s.e.}(\hat{\beta}) = \left\{\frac{\text{var}(\log(\hat{\sigma}))}{\hat{\sigma}^2}\right\}^{0.5}$
$\hat{\theta} = -\frac{\hat{\alpha}}{\hat{\sigma}}$	$\text{s.e.}(\hat{\theta}) = \left\{\left(\frac{\hat{\alpha}}{\hat{\sigma}}\right)^2 \left(\text{var}\left(\log\left(\hat{\sigma}\right)\right) + \frac{\text{var}(\hat{\alpha})}{\hat{\alpha}^2}\right)\right\}^{0.5}$
`gamlss`	
$\hat{\beta} = \exp(\hat{\sigma})$	$\text{s.e.}(\hat{\beta}) = \left\{(\exp(\hat{\sigma}))^2 \text{var}(\hat{\sigma})\right\}^{0.5}$
$\hat{\theta} = -\hat{\alpha}\exp(\hat{\sigma})$	$\text{s.e.}(\hat{\theta}) = \left\{\hat{\alpha}^2(\exp(\hat{\sigma}))^2 \left(\frac{\text{var}(\hat{\alpha})}{\hat{\alpha}^2} + \text{var}(\hat{\sigma})\right)\right\}^{0.5}$

In particular, $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_P)^T$ are the estimated regression coefficients, $\hat{\sigma}$ and $\frac{1}{\exp(\hat{\sigma})}$ are the estimated scale parameters obtained with the `survreg` and `gamlss` function, respectively. Additionally, in the `gamlss` environment when a mixed

model is considered, the variance of the random effects, i.e. σ_u^2, can be obtained as $\sigma_u^2 = \phi_u^2(\exp(\hat{\sigma}))^2$, where ϕ_u^2 is the variance of the random effects obtained by fitting a linear mixed regression model as in Equation 3.9 via a continuous Weibull distribution. Thus, to obtain the t-value one can simply compute the ratio between the estimated parameter and its respective standard error. Lastly, the probability associated to the t-value, i.e. $\Pr(\text{t-value} > t)$, can be computed as twice the output of the `pt` function which returns the distribution function of a t-distribution with (n-P-1) degrees of freedom.

3.3.3 Interpretation of the regression parameters

There is no closed form of the moments of the Discrete Weibull distribution. Nevertheless, from the estimated model we can obtain the fitted values of the conditional distribution with respect to the mean as in Equation 3.3 by numerical approximation on a truncated support of the moments of the Discrete Weibull. In `R` software this can be done with the function `Edweibull` available in the `DiscreteWeibull` package [13]. Anyway, given the usual skewed nature of the count data, an approach with regards to the conditional median may be more appropriate. Hence, substituting the formulation of the parameter $q(x) = e^{-e^{x\theta}}$ with respect to the median $\mu^{(0.5)}$ presented in Equation 3.6, leads to

$$\log(\mu^{(0.5)} + 1) \approx \frac{1}{\beta}\log(\log(2)) - \frac{1}{\beta}x\theta.$$

Thus, like for any conditional distribution which is assumed to belong to an exponential family where the parameters are linked to the mean, here the regression parameter θ can be interpreted with respect to the logarithm of the median. Specifically,

$$\frac{1}{\beta}(\log(\log(2)) - \theta_0) \tag{3.14}$$

is related to the conditional median when all the remaining covariates are set to zero, while for p^{th} covariate,

$$-\frac{\theta_p}{\beta} \tag{3.15}$$

can be related to the change in the median of the response corresponding to a one unit change of x_p while keeping all the other covariate constant.

3.4 Model selection and diagnostic

Within a likelihood based approach, we can assess the fit of our parametric model by its global deviance defined as

$$\text{GDEV} = -2\,l(y, x; \theta, \vartheta),$$

or simply by means of comparison of its log-likelihood $l(y, x; \theta, \vartheta)$. Thus, for comparing non-nested models we can use the generalised Akaike information criterion (GAIC), defined as $\text{GAIC} = \text{GDEV} + (\kappa\,\text{df})$, which includes a penalty κ for each effective degree of freedom (i.e., the number of free parameters) used in the model. To compare existing parametric approaches with our approach, we employ the special case of the GAIC corresponding to $\kappa = 2$ which leads to the Akaike information criterion (AIC) [6]

$$\text{AIC} = \text{GDEV} + (2\,\text{df}).$$

The best model is the one with the lowest AIC value.

Moreover, the appropriateness of the selected model can be assessed with a residual analysis. Given that the response is discrete, the analysis will be based on the randomised quantile residuals [33]. In particular, let

$$r_i = \Phi^{-1}(u_i)$$

for $i = 1, \dots, n$, and where Φ^{-1} is the inverse cumulative distribution function of a standard Normal variable, and u_i is a realisation from a Uniform random variable on the interval

$$[u_1; u_2] \approx [F(y_i - 1; q(x), \beta), F(y_i; q(x), \beta)].$$

The main advantage of the normalised randomised quantile residuals is that, whatever the distribution of the response variable, their true values r_i, for $i = 1, \dots, n$, always have a standard Normal distribution if the model assumption is correct. Since checking the normality assumption is well established within the statistical literature, e.g. using a qq-plot, the randomised normalised quantile residuals provide an easy way to check the adequacy of the fitted model. The randomisation of these quantile residuals is also appropriate for interval censored response variables.

In addition to the residual analysis, it is informative also to check whether the data shows any under- or over-dispersion relative to the specified Discrete Weibull conditional distribution. In the case of good fitting, we would expect the ratio of observed and theoretical variance in equation Equation 3.7 to be close to 1 for each x. In order to check for this, we produce a variance ratio plot whereby we split the response values into a number of groups of similar size, based on the percentiles of the fitted values from the specified distribution. Then the observed variance is computed within each group, while the theoretical variance from the model is averaged within each group. If the model is well specified, we would expect these values to be close to 1.

3.5 Simulation study

In this section we consider a number of simulations to asses the performance of our novel regression approach via a Discrete Weibull distribution.

3.5.1 Computational efficiency of `gamlss` and `survreg` implementations

Recalling the linear regression model in Equation 3.8, we simulate n=3,000 realisations from a Discrete Weibull with $q(x)$ and β parameters, and one covariate only. This leads to

$$\begin{aligned} \log(-\log(q(x))) &= \theta_0 + \theta_1 x \\ \log(\beta) &= \vartheta_0. \end{aligned} \tag{3.16}$$

where θ_0=-6.7, θ_1=0.9, $x \sim$ Uniform$(-1, 1)$, ϑ_0=0.7, thus $\beta \approx$2. Table 3.2 shows the parameter estimates for the linear regression model in Equation 3.16. The comparison is between the newly implemented functions `dw.gamlss` and `dw.survreg`, and the existing `dw.reg` function available in the earlier version of the `R` package `DWreg` [107], which does not employ the link with the continuous Weibull interval censored distribution. Specifically, exploiting the parametrization presented in subsection 3.3.2, the `dw.gamlss` function calls the function `gamlss` in the `R` package `gamlss`, while the `dw.survreg` calls the function `survreg` in the `R` package `survreg`. It is clear how the three functions return very similar estimates and standard errors. However, the `dw.reg` function has a higher computational cost. Using the `R` function `system.time` available in the `R` package `base`, in Table 3.3 we show a comparison of the CPU time needed to compute

TABLE 3.2: Parameter estimates for the linear regression model in Equation 3.16 via the R functions `dw.gamlss`, `dw.survreg`, and `dw.reg`.

	`dw.gamlss`	`dw.survreg`	`dw.reg`
(Intercept)	-6.732***	-6.732***	-6.732***
	(0.103)	(0.103)	(0.109)
x	0.849***	0.849***	0.849***
	(0.006)	(0.006)	(0.006)
β	2.022***	2.043***	2.043***
	(0.029)	(0.029)	(0.029)

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

a linear Discrete Weibull regression model on the datasets `rwm` available in the R package `COUNT` which includes 27,326 observations.

TABLE 3.3: CPU time performance comparison between the R functions `dw.survreg`, `dw.gamlss`, and the old version of `dw.reg` on estimating a linear Discrete Weibull regression model on the `rwm` data which contains n=27,326 observations.

Function	`dw.survreg`	`dw.gamlss`	`dw.reg`
CPU time	0.34	4.79	17.82

3.5.2 Linear mixed regression simulated data

Recalling the linear mixed regression model in Equation 3.9, we simulate data from a random intercept and a random slope model. We evaluate the estimation of parameters and describe a method for computing their standard errors.

3.5.2.1 Random intercept model

We consider a 2-level random intercept model where the level-1 observation $i = 1, \dots, n_j$ is nested in the level-2 group $j = 1, \dots, J$, and there is one covariate x. This leads to

$$\begin{aligned} \log(-\log(q(x,u)) &= \theta_{0j} + \theta_{1j} x_{ij} \\ \theta_{0j} &= \gamma_{00} + u_{0j} \\ \theta_{1j} &= \gamma_{10} \\ \log(\beta) &= \vartheta_{0j} \\ \vartheta_{0j} &= \alpha_{00}, \end{aligned}$$

where $u_{0j} \sim \text{Normal}(0, \sigma_0^2)$. This can be rewritten in full terms as

$$\begin{aligned} \log(-\log(q(x,u)) &= \gamma_{00} + \gamma_{10} x_{ij} + (u_{0j}) \\ \log(\beta) &= \alpha_{00}. \end{aligned} \tag{3.17}$$

Thus, we define $x_{.j} \sim \text{Uniform}(-1, 1)$, $\gamma_{00} = -3.9$, $\gamma_{10} = 0.7$, $\alpha_{00} = 0.7$ so that $\beta \approx 2$, and $q(x)$ varies between 0.9 and 0.99. Moreover, we assume equal sample size in each group, i.e. $n_j = 100$, and we consider $J = 15$ groups. The random effects u_{0j} are assumed i.i.d. as $u_{0j} \sim \text{Normal}(0, \sigma_0^2)$, where we set $\sigma_0^2 = 0.4$. The bar plot of the response variable simulated under

these values, and the box-plot of the response by group can be visualised in Figure 3.4. Using the `gamlss` implementation, we obtain the parameter estimates $\hat{\gamma}_{00} = -4.08$, $\hat{\gamma}_{10} = 0.72$, $\hat{\beta} = 2.02$, and $\hat{\sigma}_0^2 = 0.42$. The AIC value of the random intercept model via a Discrete Weibull distribution is 5,250.88. As a comparison, we fit the same data via a Poisson and a Negative Binomial distribution. The AIC values of these models are 5,660.33 and 5,273.25 for the Poisson and Negative Binomial model, respectively.

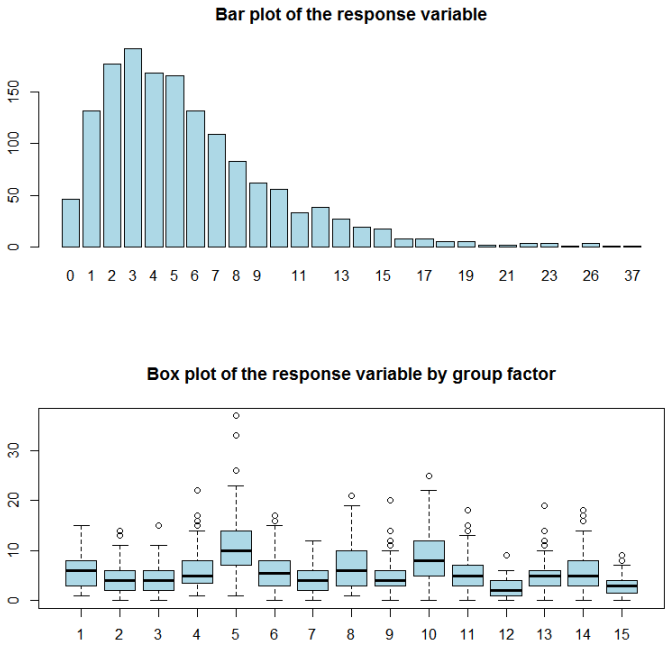


FIGURE 3.4: Bar plot and box-plot by group of a random intercept multilevel Discrete Weibull model with parameters $\beta \approx 2$, and $q(x) \in [.9, .99]$.

3.5.2.2 Random slopes model

We consider a 2-level random slopes model for the individual $i = 1, \ldots, n_j$ within the cluster $j = 1, \ldots, J$ and one covariate x. This leads to

$$\begin{aligned} \log(-\log(q(x,u)) &= \theta_{0j} + \theta_{1j} x_{ij} \\ \theta_{0j} &= \gamma_{00} + u_{0j} \\ \theta_{1j} &= \gamma_{10} + u_{1j} \\ \log(\beta) &= \vartheta_{0j} \\ \vartheta_{0j} &= \alpha_{00} \end{aligned}$$

where $u_{0j} \sim \text{Normal}(0, \sigma_0^2)$, and $u_{1j} \sim \text{Normal}(0, \sigma_1^2)$. This can be rewritten in full terms as

$$\begin{aligned} \log(-\log(q(x,u)) &= \gamma_{00} + \gamma_{10} x_{ij} + (u_{0j} + u_{1j} x_{ij}) \\ \log(\beta) &= \alpha_{00} \end{aligned} \tag{3.18}$$

where $x_{.j} \sim \text{Uniform}(-1, 1)$, $\gamma_{00} = -5.4$, $\gamma_{10} = 0.9$, $\alpha_{00} = 0.7$ so that $\beta \approx 2$, and $q(x)$ varies between 0.75 and 0.99. Moreover, we assume equal sample size in each group, i.e. $n_j = 100$, and we consider $J = 15$ groups. The random effects u_{ij} are assumed i.i.d. as $u_{ij} \sim$ multivariate Normal $\left([0 \quad 0]^T, \Sigma^2\right)$, and we set $\Sigma^2 = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$. The bar plot of the response variable simulated under these values, and the box-plot of the response by group can be visualised in Figure 3.5.

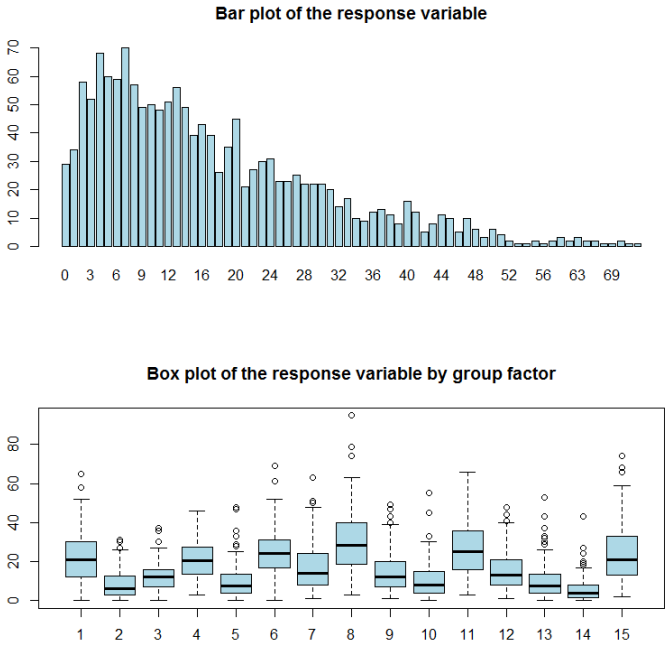


FIGURE 3.5: Bar plot and box-plot by group of a random slope multilevel Discrete Weibull model with parameters $\beta \approx 2$, and $q(x) \in [.75, .99]$.

We then employ the simulated dataset in a regression model as described in Equation 3.18. This leads to the parameter estimates $\hat{\gamma}_{00} = -5.437$, $\hat{\gamma}_{10} = 0.948$, $\hat{\beta} = 2.056$, and $\hat{\Sigma}^2 = \begin{bmatrix} 0.895 & 0.326 \\ 0.326 & 0.934 \end{bmatrix}$. The AIC value of the random slopes model via a Discrete Weibull distribution is 9,729.39. As a comparison, we fit the same data via a Poisson and Negative Binomial distribution. The AIC values of these models are 12,247.98 and 9,751.71 for the Poisson and Negative Binomial model, respectively.

3.5.2.3 Parametric bootstrap estimation of the standard errors

Bootstrapping is a re-sampling method for statistical inference [34]. It consists in repeatedly drawing random samples from the original sample, with replacement and with the same size of the original sample. Thus, the sampling distribution of the bootstrap estimates of a parameter of interest is obtained from the pool of bootstrap re-samples, as well as the biased-corrected estimate, standard error, and confidence interval of the parameter. We apply this procedure in order to obtain the standard errors of the random effects of our Discrete Weibull mixed regression model.

Random intercept model We use the parameter estimates from the model in Equation 3.17, i.e. $\hat{\gamma}_{00}$, $\hat{\gamma}_{10}$, $\hat{\beta}$, and $\hat{\sigma}_0^2$, to compute the new response variable $Y^*|X \sim$ Discrete Weibull$(\hat{q}(x,u),\hat{\beta})$, and we refit the model via the formulation presented in Equation 3.17. This is repeated $b = 1000$ times leading to the bootstrap estimates $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, $\hat{\beta}_{(b)}$, and the variance of the random effects $\hat{\sigma}^2_{0(b)}$. Thus, we can compute the standard error as the standard deviation of the empirical distribution of the parameter estimates. This leads to the results in Table 3.4. Moreover, for $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, $\hat{\beta}$, and $\hat{\sigma}^2_{0(b)}$ we compute a coverage measure based on the 95% confidence interval. We iterate the bootstrap procedure $k = 200$ times in order to compute these values as the proportion of instances in which the true parameter γ_{00}, γ_{10}, β, and σ_0^2 was found in its respective 95% bootstrap confidence interval. The resulting percentage coverage is expected to be close to the nominal confidence of the interval estimate. This leads to a coverage values of 95%, 96%, 95%, 97% for the parameters γ_{00}, γ_{10}, β,

TABLE 3.4: Parameter estimates for the simulated random intercept Discrete Weibull model with standard errors in brackets obtained via a parametric bootstrap approach.

Fixed effects		Random part	
(Intercept)	-4.077***	σ_0^2	0.416***
	(0.158)		(0.059)
x	0.717***		
	(0.027)		
β	2.018***		
	(0.025)		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

and σ_0^2, respectively. The plot of the bootstrapped estimates is presented in Figure 3.6 and shows distributions centred around the true values.

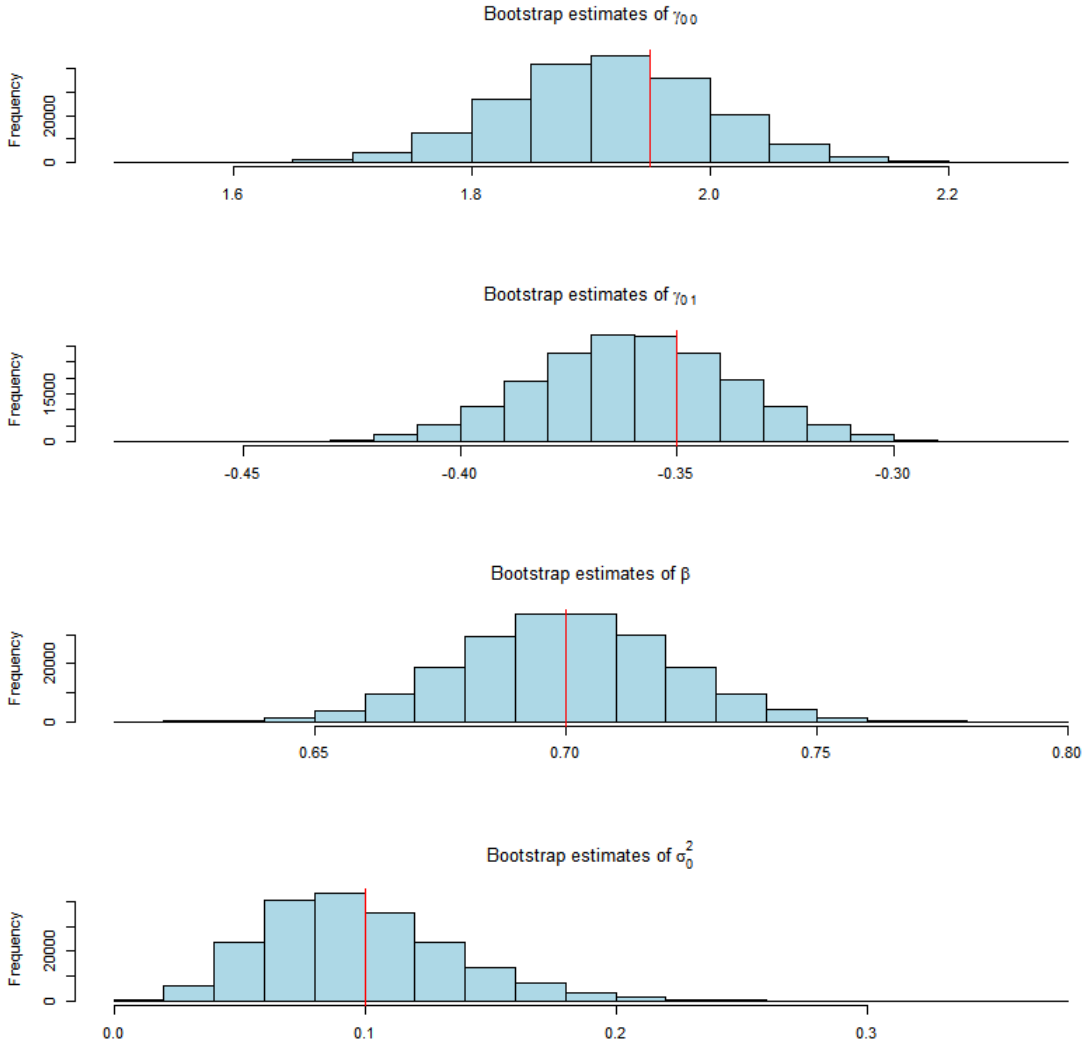


FIGURE 3.6: True parameter (red line) and distribution of the bootstrapped estimates $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, $\hat{\beta}_{(b)}$, and $\hat{\sigma}^2_{0(b)}$ of a simulated random intercept multilevel Discrete Weibull model obtained over $k = 200$ iterations of $b = 1000$ bootstrap replications.

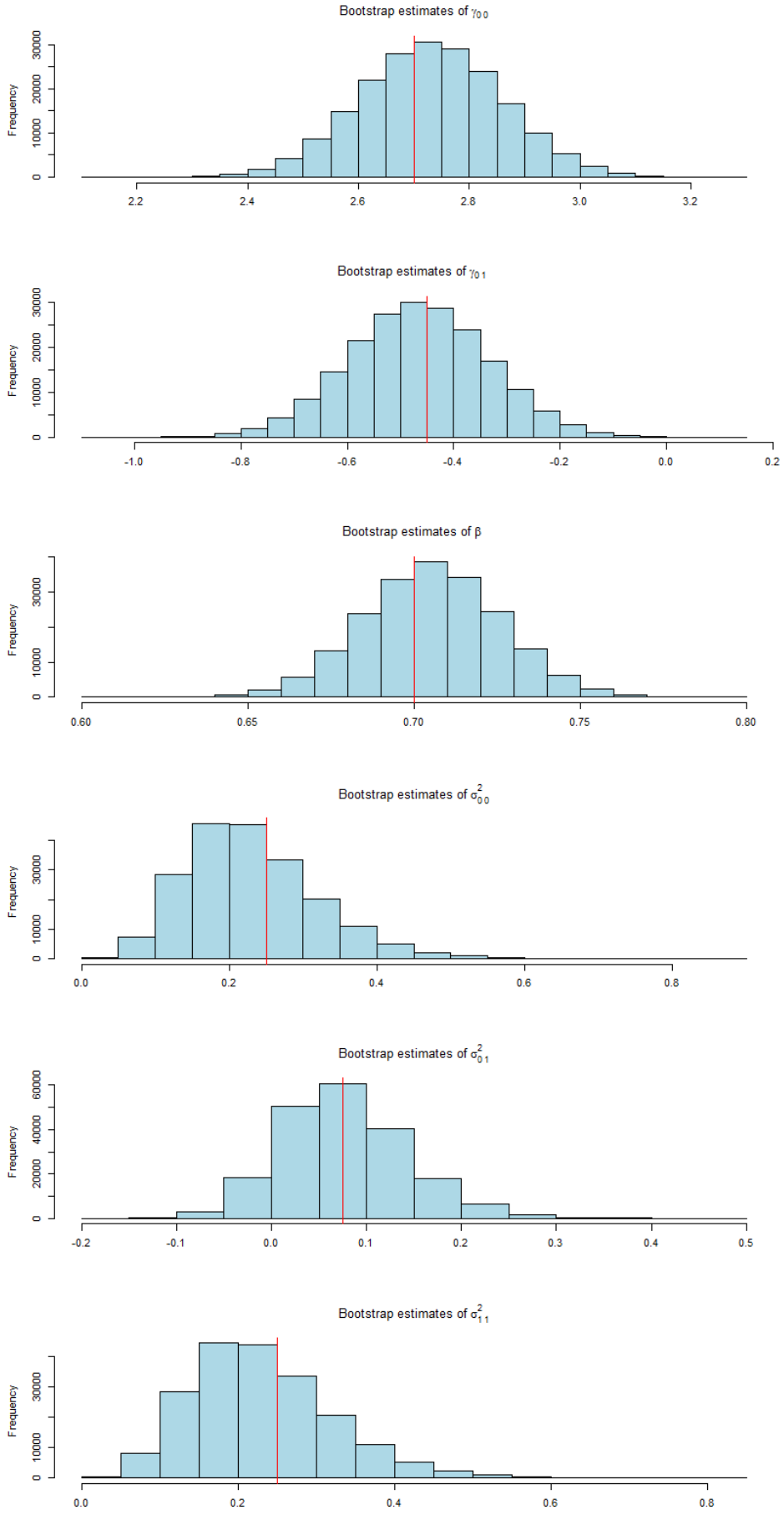


FIGURE 3.7: True parameter (red line) and distribution of the bootstrapped estimates $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, $\hat{\beta}_{(b)}$, and $\hat{\sigma}^2_{00(b)}$, $\hat{\sigma}^2_{01(b)} = \hat{\sigma}^2_{10(b)}$, $\hat{\sigma}^2_{11(b)}$ of a simulated random slope multilevel Discrete Weibull model obtained over $k = 200$ iterations of $b = 1000$ bootstrap replications.

Random slopes model We use the parameter estimates from the model in Equation 3.18 to compute the new response variable $Y^*|X \sim$ Discrete Weibull$(\hat{q}(x,u),\hat{\beta})$. Next, we refit the model via the random slopes model formulation. This is repeated $b = 1000$ times, and it leads to the bootstrap estimates $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, $\hat{\beta}_{(b)}$, and to the variance-covariance matrix of the random effects $\hat{\Sigma}^2_{(b)}$. Thus, we compute the standard error as the standard deviation of the empirical distribution of the parameters, which leads to the estimates in Table 3.5.

TABLE 3.5: Parameter estimates for the simulated random slope Discrete Weibull model with standard errors in brackets obtained via a parametric bootstrap approach.

Fixed effects		Random part	
(Intercept)	-5.469***	σ^2_{00}	0.985***
	(0.132)		(0.09)
x	0.931***	σ^2_{01}	0.326***
	(0.135)		(0.066)
β	2.056***	σ^2_{11}	0.993***
	(0.02)		(0.09)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Then, we iterate the procedure $k = 200$ times in order to compute the coverage value as the 95% confidence interval of the parameters $\hat{\gamma}_{00(b)}$, $\hat{\gamma}_{10(b)}$, and $\hat{\sigma}^2_{00(b)}$, $\hat{\sigma}^2_{01(b)} = \hat{\sigma}^2_{10(b)}$, $\hat{\sigma}^2_{11(b)}$. This leads to 96%, 94%, 95% coverage values for the parameters γ_{00}, γ_{10}, and β respectively, and 96%, 95%, and 96% coverage values for the variance-covariance matrix of the random effects $\hat{\sigma}^2_{00}$, $\hat{\sigma}^2_{01} = \hat{\sigma}^2_{10}$, $\hat{\sigma}^2_{11}$, respectively. The plot of the bootstrapped estimates via the `gamlss` parametrization is shown in Figure 3.7.

3.5.3 Excess zeros regression simulated data

We now consider the case when the data are inflated by an excess of zeros. Thus, we simulate n=2,000 realisations from a mixture model combining a constant logit to model the zeros of the response, and a count model via the Discrete Weibull distribution with parameters $q(x)$ and β. Specifically, this can be written as

$$Pr(Y|X) = \begin{cases} \pi + (1-\pi)(1-q(x)) & \text{for } y = 0 \\ (1-\pi) \begin{cases} \log(-\log(q(x))) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \\ \log(\beta) = \vartheta_0 \end{cases} & \text{for } y = 1,2,3,\ldots, \end{cases} \tag{3.19}$$

where $\pi = 0.6$, $x_1 \sim$ Uniform$(-1,1)$, $x_2 \sim$ Uniform$(0,1)$, $x_3 \sim$ Normal$(0,1)$, $\theta_0 = -3.5$, $\theta_1 = -2.4$, $\theta_2 = 0.8$, $\theta_3 = -0.3$, and $\vartheta_0 = 0.7$. This leads to $\beta \approx 2$, and $q(x)$ which varies between 0.38 and 0.99. The percentage of zeros in the data is 63.5%, and the dependent variable has a dispersion index of 7.52. The bar plot of the response variable simulated under these values can be visualised in Figure 3.8.

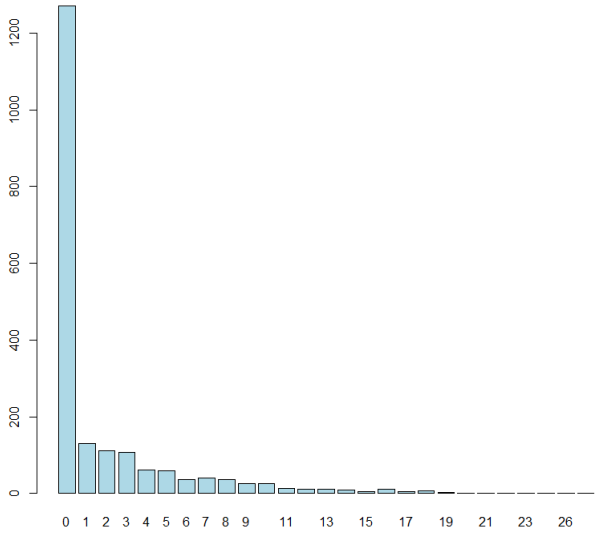


FIGURE 3.8: Bar plot of the simulated zero inflated Discrete Weibull model with $\beta \approx 2$, and $q(x) \in [.38, .99]$.

Thus, we fit a zero inflated Discrete Weibull and a hurdle Discrete Weibull model with parameters $q(x)$ and β as in Equation 3.10 and Equation 3.12, respectively. We compare these models with a zero inflated and a hurdle model via a Poisson distribution as in Equation 1.1 and Equation 1.3, respectively, and with a zero inflated and a hurdle model via a Negative binomial distribution as presented in Equation 1.2 and Equation 1.4, respectively. To perform the Poisson and Negative Binomial zero inflated and hurdle regression models we use the `zeroinfl` and `hurdle` functions available in the `R` package `pscl` [55]. The parameter estimates are presented in Table 3.6. Specifically, we report the parametrisation to the logarithm of the mean for the Poisson and Negative Binomial model, while for the count model via the Discrete Weibull we employ the parametrisation to the logarithm of the median presented in Equation 3.14 and Equation 3.14, for the intercept and the three covariates, respectively.

TABLE 3.6: Parameter estimates and AIC for the simulated zero-excessive Discrete Weibull model in Equation 3.19 fitted by using different parametric zero inflated and hurdle models.

	ZI PO	ZI NB	ZI DW	hurdle PO	hurdle NB	hurdle DW
(Intercept)	1.483***	1.451***	1.565***	1.481***	1.444***	1.565***
	(0.037)	(0.05)	(0.176)	(0.037)	(0.051)	(0.177)
x1	1.33***	1.415***	1.181***	1.329***	1.416***	1.175***
	(0.038)	(0.052)	(0.11)	(0.039)	(0.053)	(0.11)
x2	-0.398***	-0.428***	-0.351***	-0.393***	-0.417***	-0.346***
	(0.061)	(0.087)	(0.136)	(0.061)	(0.087)	(0.137)
x3	0.154***	0.162***	0.132***	0.154***	0.163***	0.132***
	(0.019)	(0.026)	(0.039)	(0.019)	(0.026)	(0.04)
other	-	σ=1.954***	β=2.099***	-	σ=1.95***	β=2.103***
	-	(0.149)	(0.079)	-	(0.149)	(0.079)
AIC	5574.748	5444.528	5417.350	5580.753	5450.210	5423.893
logLik	-2779.374	-2713.264	-2699.675	-2782.377	-2716.105	-2702.946

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The AIC values reported for both the model formulation, i.e. zero inflated and hurdle, point to the choice of a Discrete Weibull model as the best fitting one, followed by the Negative Binomial model and, lastly, by the Poisson model. In particular, the zero inflated model formulation fits this data slightly better than the hurdle regression. As expected given the choice of a constant π to model the zeros, we note that the logit coefficient of x_1, x_2, and x_3 are not significant, thus they are omitted.

3.6 Real data study

We now consider a number of real data examples to illustrate our parametric approach via a Discrete Weibull distribution. Specifically, we will consider both cases of over- and under- dispersion and excessive zeros data.

3.6.1 Over-dispersed data

Length of stay in hospital The results in chapter 2 have shown a positive effect of the P4P program on the hospital effectiveness in the Italian Lombardy region, but what about their efficiency? In this analysis we aim to study the effect of the P4P on three patient health conditions, namely coronary artery bypass graft surgery (CABG), percutaneous transluminal coronary angioplasty (PTCA) and hip replacement (HIP). As outcome, we consider the in-hospital patient's length of stay which is a commonly used indicator of the quality of care and planning capacity within a hospital [10, 20, 49, 50] since it is a proxy of the expenditure of each hospitalisation in a DRG-based payment system [97]. Thus, a reduction in this measures provides a reduction in the hospital costs for the same reimbursement and this drives an increment of the hospital efficiency. In this sense, our analysis can be seen as an evaluation of the relationship between hospital efficiency and the quality of care provided. The data used are gathered from the Lombardy healthcare information system regarding patients admitted to either public or private hospitals during the year 2014. For the analysis, we subset the data by excluding patients living outside the region and patients younger than two years old. Thus, for the three health conditions described above, we used a total of 23,709 hospitalisations within 110 hospitals of the Lombardy region, of which 3,851 hospitalisations were for CABG, 7,083 for HIP, and 12,775 for PTCA, respectively. The average length of stay for patient admitted for CABG is approximately 15 days, for HIP is approximately 8 days, and for PTCA is approximately 10 days. The evaluation of the P4P impact in terms of length of stay will be described including patients' demographic characteristics while considering the severity of the health condition. Specifically, we consider the gender and age of the patients, the comorbidity index measured as in [38], and a factor variable with categories the three procedures or patient-reported health conditions. The length of stay of the hospitalisation is measured in days and obtained as a difference between the discharge and the in-hospital admission date. This variable is over-dispersed, with a mean of 10.26, a range of [0;144], and a dispersion value of 5.45. Our empirical approach is to estimate a multilevel model that recognises the clustering of patients within providers. Specifically, we estimate multilevel models with provider-specific intercepts [81, 92], where the patients are the level-1 observations and the hospitals represent the level-2 units. We compare our approach with models of the same complexity assuming a Poisson, a Poisson-inverse Gaussian, a COM-Poisson, and a Negative Binomial distribution, as these are the most widespread parametric approaches for modelling over-dispersed count data. The Poisson, the Negative Binomial, and the Poisson-inverse Gaussian model are implemented via the function `gamlss` in the `R` package `gamlss` [83], and the COM-Poisson is implemented via the function `HLfit` in the `R` package `spaMM` [85]. Figure 3.9 shows the empirical distribution of the response variable.

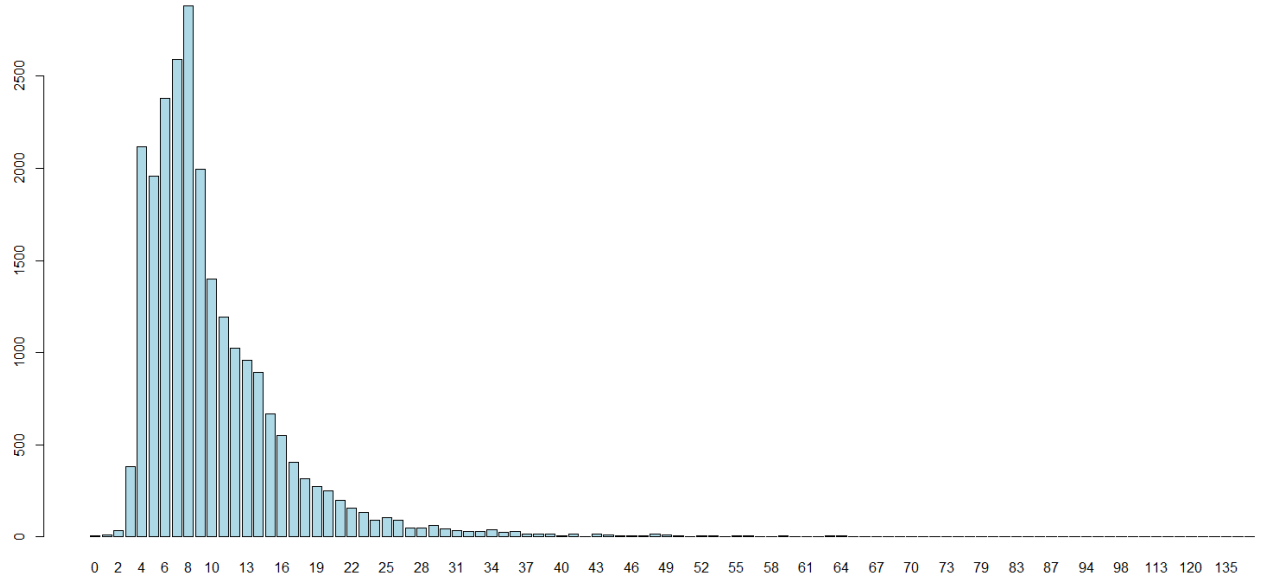


FIGURE 3.9: Bar plot of hospital length of stay measured in days.

For the patient p (with $p = 1, \dots, P_h$) hospitalised within hospital h (with $h = 1, \dots, H$), the linear mixed Discrete Weibull model can be written as

$$\begin{aligned} \log(-\log(q(x,u)) = \theta_0 + \theta_1 \text{female}_{ph} + \theta_2 \text{age}_{ph} + \theta_3 \text{comorbidity1}_{ph} + \theta_4 \text{comorbidity2}_{ph} + \\ \theta_5 \text{comorbidity3}_{ph} + \theta_6 \text{procHIP}_{ph} + \theta_7 \text{procCABG}_{ph} + u_h \\ \log(\beta) = \vartheta_0, \end{aligned} \tag{3.20}$$

where $\theta = (\theta_0, \theta_1, \dots, \theta_7)^T$ is the vector of coefficients for the patients level covariates, and u_h is the random effect for hospital h. Table 3.7 reports the parameter estimates and the AIC values for this model and the same specification model via the distributions described above. The AIC values describe the good performance of the COM-Poisson, Negative Binomial and Discrete Weibull models. The parameters of the Poisson, COM-Poisson, Negative Binomial and Poisson-Inverse Gaussian model are linked to the logarithm of their expected mean, while the Discrete Weibull model is parametrised with respect to the logarithm of the median as presented in Equation 3.14 and Equation 3.15. For the random part of the Discrete Weibull mixed effects model the standard error is obtained over 1000 bootstrap replications. We consider the fixed part of the model presented in Table 3.7 to investigate how the patient-level factors and the diagnosis procedure explain the variations in the hospital length of stay. The effects of the age and the presence of comorbidities are associated with an increment of the length of stay, while patients admitted for HIP generally have a shorter stay than patients admitted for PTCA, while patients admitted for CABG generally have a longer stay than patients admitted for PTCA. The sex of the patient is not statistically significant. We now consider the random part of the model, and to offer a visual comparison of the effects across hospitals, the intercept estimates of each hospital are plot in Figure 3.10. Specifically, we consider the variation of each hospital with respect to the red line, i.e. the fixed intercept estimate. The blue lines represent the 25% and 75% quantiles of the distribution of the random effects, respectively. Thus, assuming that all hospitals aim to make efficiency savings, we interpret these effects as a measure of the hospitals' performance and we identify the hospitals which have been more successful in terms of shorter in-hospital stay, i.e. green dots, after taking into account the characteristics of the patients being treated and their health condition.

TABLE 3.7: Parameter estimates and AIC values for the mixed effects models with hospitals random effects for the length of stay data.

	PO	CMP	NB	PIG	DW
Fixed part					
(Intercept)	1.62***	1.486	1.535***	1.109***	1.327***
	(0.054)	(0.034)	(0.059)	(0.155)	(0.054)
female	0.003	0.005	0.002	0.005	0.003
	(0.011)	(0.004)	(0.012)	(0.034)	(0.011)
age	0.009***	0.007	0.009***	0.008***	0.009***
	(0.001)	(0)	(0.001)	(0.002)	(0.001)
comorbidity1	0.207***	0.18	0.2***	0.156***	0.207***
	(0.019)	(0.007)	(0.02)	(0.045)	(0.019)
comorbidity2	0.302***	0.268	0.289***	0.211**	0.302***
	(0.033)	(0.011)	(0.036)	(0.069)	(0.033)
comorbidity3	0.251***	0.237	0.251***	0.22	0.251***
	(0.071)	(0.021)	(0.076)	(0.174)	(0.071)
procHIP	-0.11***	-0.112	-0.113***	-0.103*	-0.11***
	(0.012)	(0.005)	(0.013)	(0.043)	(0.012)
procCABG	0.415***	0.352	0.412***	0.372***	0.415***
	(0.017)	(0.007)	(0.018)	(0.043)	(0.017)
other	-	σ=-2.515	σ=0.657***	σ=1.142***	β=1.25***
	-	(0.136)	(0.102)	(0.128)	(0.09)
Random part					
var(u_h)	0.093	0.081	0.094	0.128	0.129
	-	-	-	-	(0.009)
AIC	152120.4	149371.7	149285.3	150375.8	149210.4
logLik	-75946.7	-74927.34	-74528.5	-75106.5	-74491.7

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

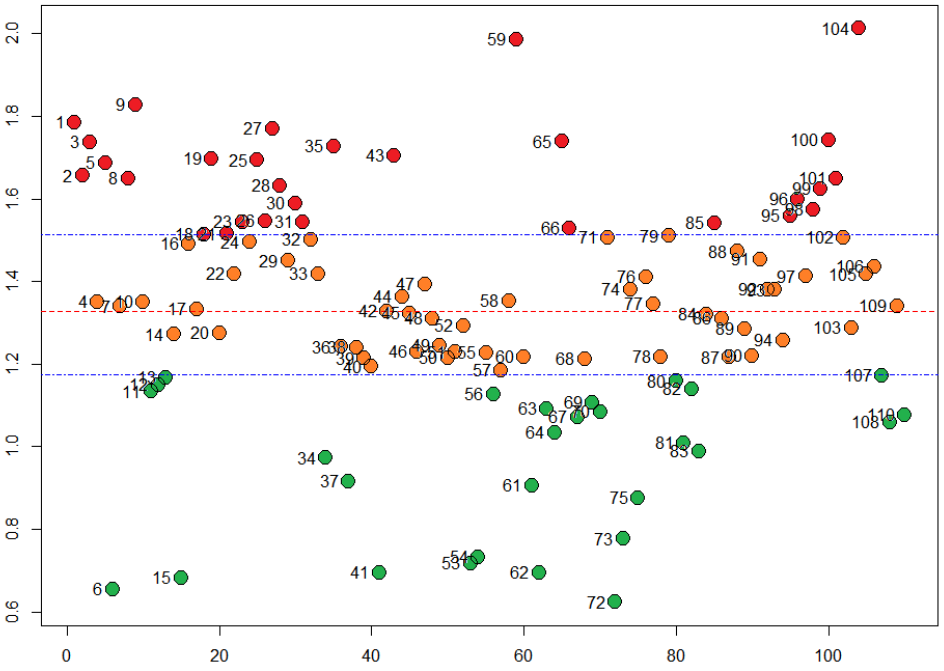


FIGURE 3.10: Parameter estimates for the random part of the mixed Discrete Weibull model on the length of stay data. The hospitals allocated below the expected median value of the response (red line) show good performances in terms of efficiency.

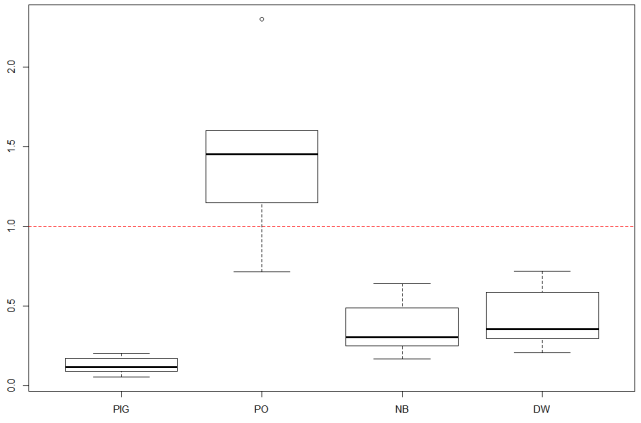


FIGURE 3.11: Variance ratio plot for the models fitted to the hospital length of stay data.

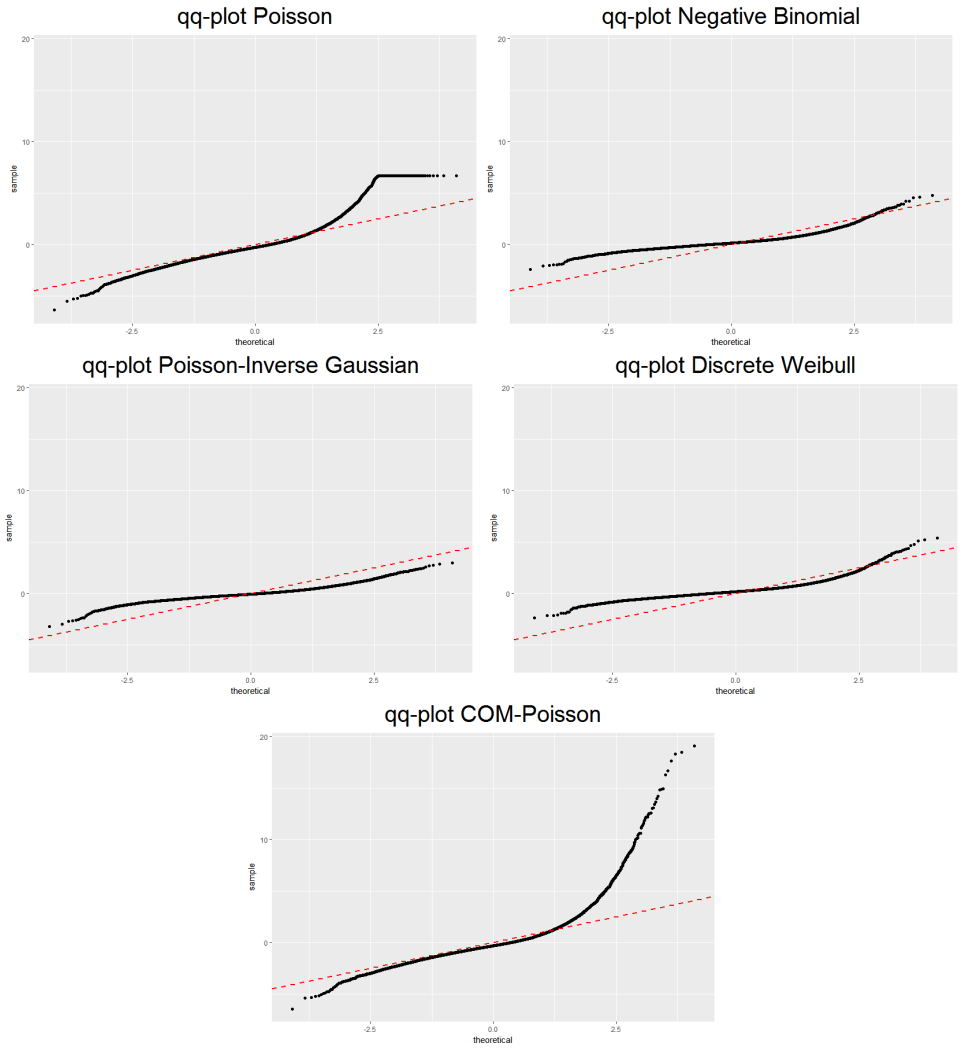


FIGURE 3.12: Diagnostic plots of the theoretical versus the sample quantiles for the analyses of the length of stay data using various regression models.

The variance ratio plot between the observed variance and the averaged theoretical conditional variance for each model

is shown in Figure 3.11. Given the complexity of the model and the not straightforward formulation of the theoretical conditional variance of the COM-Poisson, the variance ratio plot for this distribution is omitted. The plot confirms the above results pointing out the Discrete Weibull and the Negative Binomial as the best performing models being its variance ratio closest to 1. The diagnostic plots in Figure 3.12 show the normalized randomized quantile residuals for the Poisson, Negative Binomial, Poisson-inverse Gaussian, Discrete Weibull, and COM-Poisson model, respectively. We can conclude that the residuals of the Discrete Weibull, Negative Binomial and Poisson-Inverse Gaussian models in general behave better than the residuals of the other models.

3.6.2 Under-dispersed data

Apgar index In the following study we investigate the Apgar score which is an index used to assess how a baby is doing at birth [9]. Given the fact that low Apgar scores are associated with a greater risk of problems, the aim of this study is to predict the medical assistance needed, and thus the cost of the hospitalisation of the newborn. The data used are gathered from the Lombardy healthcare information system regarding 55,637 baby births in 2012 in 62 hospitals of the Lombardy region in Italy. Figure 3.13 shows the bar plot of the response variable. This variable has a range [0,10] where a score of 10 means that the baby is doing very well at birth. Moreover, the response variable has a mean of 9.789, a variance of 0.471, and a dispersion value close to 0.05, thus we are modelling highly under-dispersed data.

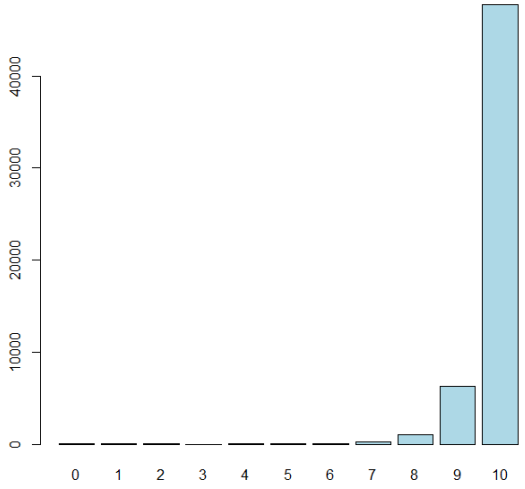


FIGURE 3.13: Bar plot of the Apgar index.

The evaluation of the cost of the hospitalisation of the newborn via the Apgar index will be described including the age of the mother and whether she delivered naturally or by caesarean section, and some physical conditions of the baby measured straight after his/her birth which are usually representative of his/her future health condition. Specifically, we consider the weight of the baby, the circumference of the baby's head, and the presence of malformations. Given the nature of the data we fit a random intercept model which considers the hospitals as the level-2 units. Specifically, for the baby b (with $b = 1, \dots, B_h$) born within hospital h (with $h = 1, \dots, H$), the linear mixed Discrete Weibull model can be written as

$$\begin{aligned} \log(-\log(q(x,u)) = \theta_0 &+ \theta_1 \text{motherage}_{bh} + \theta_2 \text{babyweight}_{bh} + \theta_3 \text{headcirc}_{bh} + \\ &\theta_4 \text{malfo2}_{bh} + \theta_5 \text{comorbidity3}_{bh} + \theta_6 \text{csection1}_{bh} + u_h \\ \log(\beta) &= \vartheta_0. \end{aligned} \tag{3.21}$$

where $\theta = (\theta_0, \theta_1, \ldots, \theta_6)^T$ is the vector of coefficients for the patients level covariates, and u_h is the random effect for hospital h. We compare the results of the Discrete Weibull model in Equation 3.21 with the same specification models obtained via a Poisson, a COM-Poisson, and a Generalised-Poisson distribution, as these are the commonly used distribution in the case of under-dispersed data.

TABLE 3.8: Parameter estimates and AIC for the mixed effect models via different distributions on the Apgar index data using random effects for the hospitals.

	PO	CMP	GPO	DW
Fixed part				
(Intercept)	2.082***	6.409.	-4.245***	2.27***
	(0.036)	(0.06)	(0.09)	(0.364)
motherage	0.001	0.001***	0.000	0.001***
	(0.001)	(0.001)	(0.001)	(0.001)
babyweight	0.001***	0.001***	0.000	0.001***
	(0.001)	(0.001)	(0.001)	(0.001)
headcirc	0.003***	0.01**	0.003	0.001***
	(0.001)	(0.002)	(0.003)	(0.001)
malfo2	0.045*	0.129*	0.045	0.02***
	(0.02)	(0.035)	(0.043)	(0.002)
csection1	-0.011***	-0.033**	-0.011	-0.006***
	(0.003)	(0.005)	(0.013)	(0.001)
other	-	σ=3.014	σ=-36.04	β=30.61***
	-	-	(-423.95)	(2.768)
Random part				
var(u_h)	0.001	0.002	0.001	0.17
	-	-	-	(0.001)
AIC	233531.8	179911.8	233442.7	76761.72
logLik	-116700	-90009.5	-116702	-38311.95

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

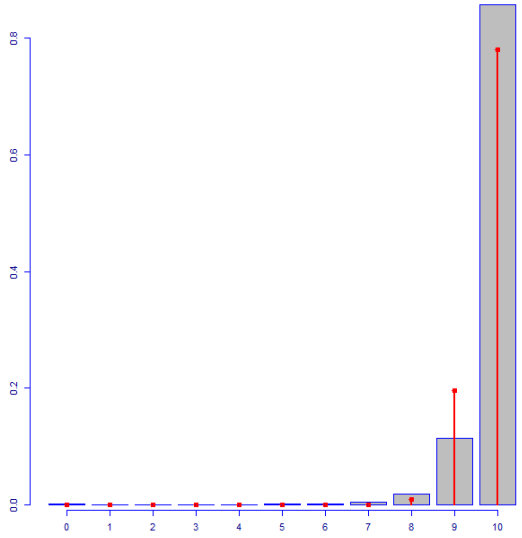


FIGURE 3.14: Observed (grey) and expected (red) frequencies for the Discrete Weibull mixed effect models on the Apgar index data.

Table 3.8 reports the parameter estimates with respect to the logarithm of the mean for the Poisson, Negative Binomial and COM-Poisson model, and with respect to the logarithm of the median for the Discrete Weibull model. The AIC and log-likelihood values of these models point out the Discrete Weibull distribution as the best fitting one. For the random part of the model the bootstrapped standard error is obtained over 1000 replications. Moreover, we measure the fit of the Discrete Weibull model by the comparison between the observed, i.e. grey bars, and expected, i.e. red lines, frequencies as shown in Figure 3.14. This is obtained by adapting to a Discrete Weibull fit the `histDist` function available in the `R` package `gamlss`.

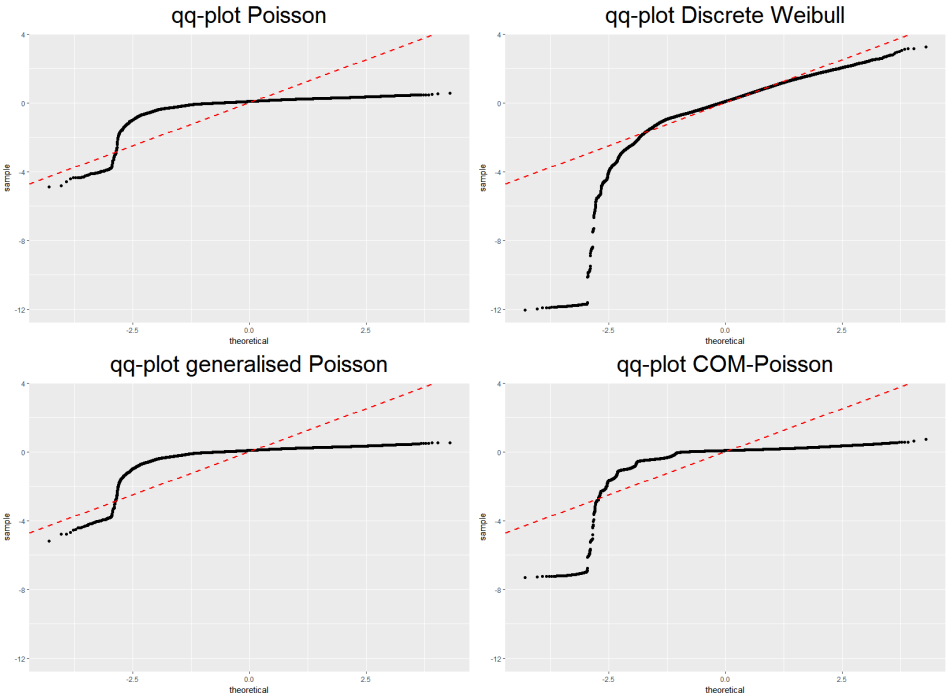


FIGURE 3.15: Diagnostic plots of the theoretical versus the sample quantiles for the the mixed effect models on the Apgar index data.

From Figure 3.15, we can conclude that the normalised quantile residuals of the Discrete Weibull model behave better than the residuals of the Poisson, COM-Poisson and Generalised Poisson model, although there is still some inaccuracy in modelling the left-tail of the distribution.

Asthma inhaler For this analysis we use the data from [45] which consists of 5,201 observations regarding the daily count of using Albuterol asthma oral inhaler for 48 children undertaking at least 30 measurements during the year. Hence, this can be seen as an example of growth models which are an important variation of multilevel models. In growth models repeated observations from an individual represent the level-1 variables, and the attributes of the individual represent the level-2 variables. In particular, the study investigates the relationship between the asthma inhaler use of each child which represents the asthma severity, and the air pollution which is recorded by four covariates: the percentage of humidity, the barometric pressure, the average daily temperature, and the morning levels of PM_{25}. The response variable has a mean of 1.27, a variance of 0.84, and a dispersion value of approximately 0.664. The observed frequencies can be visualised in Figure 3.16.

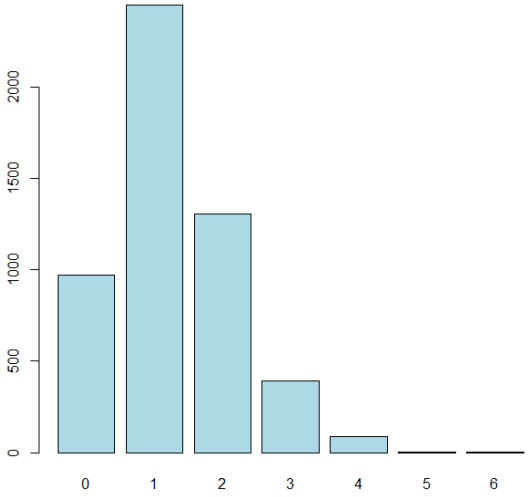


FIGURE 3.16: Bar plot of the daily count of using the asthma inhaler.

At first we fit a linear Discrete Weibull fixed effects regression model, as in Equation 3.8 to predict the response variable by including the four covariates representing the air pollution, and the child id factor. Given the under-dispersed nature of the data, we compare our approach with models of the same complexity based on a Poisson, a COM-Poisson, and a Generalised Poisson distribution. The Poisson, and the Generalised Poisson models are implemented via the function `gamlss` in the `R` package `gamlss`, while the COM-Poisson model with only fixed effects is implemented via the function `glm.comp` in the `R` package `CompGLM` [77], and the COM-Poisson mixed effects model implemented next is fitted with the `HLfit` function available in the `R` package `spaMM`. For these models, Table 3.9 shows the AIC and log-likelihood values which point to the good performances of the COM-Poisson and Discrete Weibull models.

TABLE 3.9: Comparison of the models in terms of AIC and using fixed effects only for the asthma inhaler data.

	PO	GPO	CMP	DW
AIC	13356.41	13358.43	12448.87	12446.08
logLik	-6626.204	-6626.217	-6171.435	-6170.038

The variance ratio plot in Figure 3.17 confirms the good performance of the Discrete Weibull model.

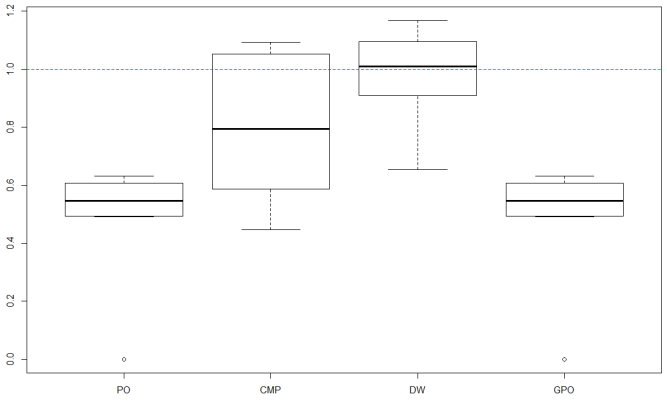


FIGURE 3.17: Variance ratio plots of four different models using fixed effects only fitted on the asthma inhaler data.

Given the structure of the data, we then fit a more appropriate random intercept multilevel model considering the child as the level-2 variable. Specifically, for the measurements m (with $m = 1, \dots, M_c$) undertaken by child c (with $c = 1, \dots, C$), the linear mixed Discrete Weibull model can be written as

$$\begin{aligned} \log(-\log(q(x,u)) &= \theta_0 + \theta_1 \text{hum.orig}_{mc} + \theta_2 \text{pres.orig}_{mc} + \\ &\quad \theta_3 \text{temp.orig}_{mc} + \theta_4 \text{pm25.orig}_{mc} + u_c \\ \log(\beta) &= \vartheta_0. \end{aligned} \tag{3.22}$$

where $\theta = (\theta_0, \theta_1, \dots, \theta_4)^T$ is the vector of coefficients for the Albuterol asthma oral inhaler measurements level covariates, and u_c is the random effect for child c.

Table 3.10 reports the parameter estimates and the AIC values of the fitted mixed models which points again to the choice of the Discrete Weibull and the COM-Poisson as the best fitting models. For the random part of the mixed effects Discrete Weibull model the bootstrapped standard error is obtained over 1000 replications.

TABLE 3.10: Comparison of the mixed effects models on the asthma inhaler data using a random effects for the children.

	PO	GPO	CMP	DW
Fixed part				
(Intercept)	-2.468	-2.421	-4.008	-0.257***
	(1.711)	(2.435)	(2.351)	(0.01)
hum.orig	-0.103	-0.104	-0.195	-0.057.
	(0.084)	(0.115)	(0.115)	(0.031)
pres.orig	4.494.	4.477	8.464	1.376***
	(2.721)	(3.875)	(3.737)	(0.185)
temp.orig	-0.188	-0.189	-0.353	-0.149*
	(0.129)	(0.172)	(0.178)	(0.062)
pm25.orig	0.021	0.021	0.039	0.005*
	(0.013)	(0.019)	(0.018)	(0.002)
other	-	σ=-36.04	σ=2.457	β=2.478***
	-	(1385.55)	-	(0.224)
Random part				
var(u_c)	0.103	0.105	0.343	0.29
	-	-	-	(0.002)
AIC	13355.83	13351.58	12445.04	12444.26
logLik	-6626.89	-6628.26	-6277.08	-6169.75

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The diagnostic plots in Figure 3.18 show the normalized randomized quantile residuals for each fitted model, confirming a good fit of the Discrete Weibull model to this data.

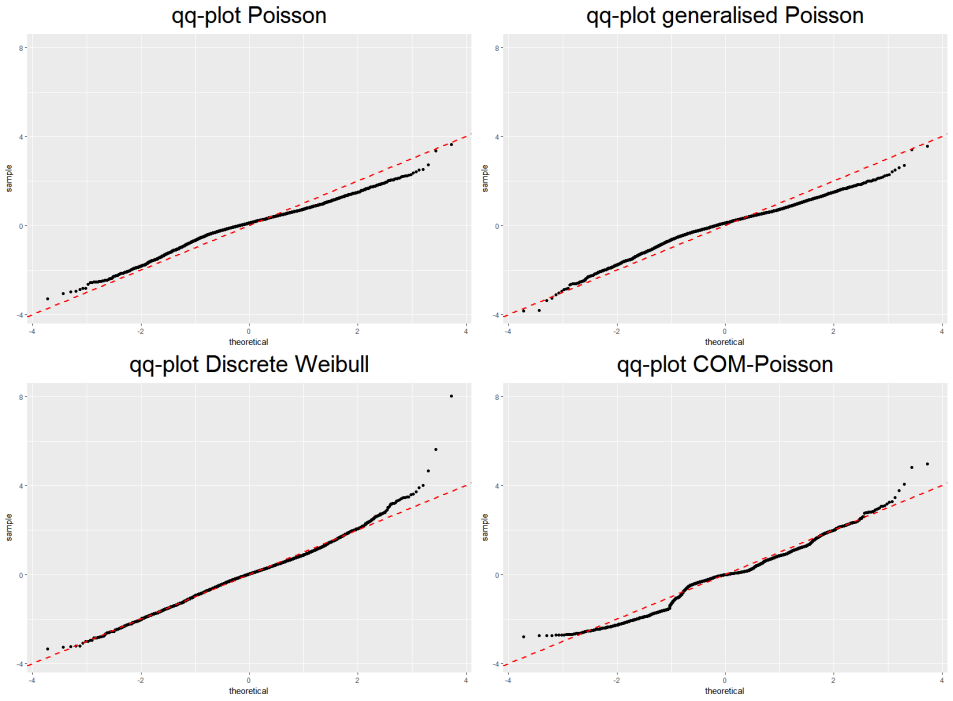


FIGURE 3.18: Diagnostic plots of the theoretical versus the sample quantiles for the analysis of the asthma inhaler data using different mixed effects regression models.

3.6.3 Excess zeros data

An excess of zeros in the data reduces the mean of the response inflating the dispersion index, thus it is crucial to consider a flexible distribution which can simultaneously account for the excess of zeros and potential over- or under- dispersion. We address this by employing Discrete Weibull zero-inflated and hurdle regression models as presented in Equation 3.10 and Equation 3.12, respectively.

Visits to physicians offices The individual number of visits to a doctor is largely used as an outcome measure of accessibility to an health service. Thus, we model the number of doctor visits considering two different examples, namely the German health registry for the year 1984, and the German socio-economic panel data.

German health registry for the year 1984

To illustrate how the Discrete Weibull handles the case of excessive zero counts, we consider the German health registry data study available in the `R` package `COUNT` under the name of `rwm1984`. This is a subset from the year 1984 of the cross-section study `rwm5yr` regarding the health information for the years immediately prior to the health reform carried in Germany. The dataset contains 3,874 observations. The number of doctor visits is regressed over the age of the patient, the gender of the patient, the working condition, the years of formal education, and the household yearly income. The response variable has approximately 42% of zeros, a mean of 3.16, a variance of 39.39, and a - possibly inflated - dispersion value of approximately 12. The observed frequencies can be shown in Figure 3.19.

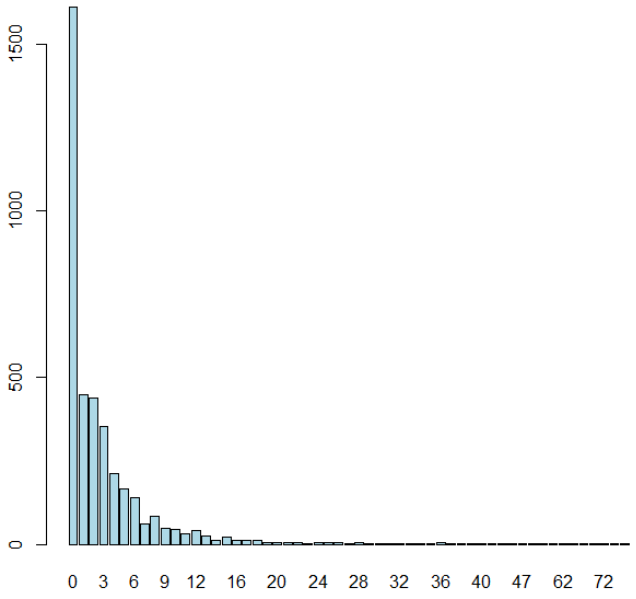


FIGURE 3.19: Bar plot of the number of doctor visit for the year 1984.

TABLE 3.11: Parameter estimates and AIC for the count part of the zero-inflated and hurdle model on the number of doctor visit for the year 1984 data.

	ZI PO	ZI NB	ZI DW	hurdle PO	hurdle NB	hurdle DW
(Intercept)	1.253***	0.536*	-0.169	1.252***	0.707**	0.026*
	(0.074)	(0.214)	(0.152)	(0.074)	(0.229)	(0.179)
age	0.013***	0.02***	0.019***	0.013***	0.018***	0.017***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)	(0.002)
female	0.156***	0.214***	0.234***	0.156***	0.206***	0.222***
	(0.019)	(0.059)	(0.043)	(0.019)	(0.061)	(0.043)
hhninc	-0.081***	-0.07***	-0.054***	-0.083***	-0.084***	-0.074***
	(0.007)	(0.016)	(0.011)	(0.008)	(0.018)	(0.012)
educ	-0.004	-0.009	-0.001	-0.003	-0.016	-0.007
	(0.005)	(0.014)	(0.01)	(0.005)	(0.015)	(0.011)
work	-0.033	-0.045	-0.099	-0.033	0.001	-0.057
	(0.044)	(0.128)	(0.091)	(0.044)	(0.139)	(0.099)
other	-	σ=-0.604***	β=0.736***	-	σ=-0.725***	β=0.721***
	-	(0.058)	(0.015)	-	(0.091)	(0.024)
AIC	24199.34	16585.74	16533.5	24195.96	16577.34	16528.92
logLik	-12087.7	-8279.87	-8253.75	-12086	-8275.67	-8251.46

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Following the model formulation for the zero-inflated and hurdle formulation presented in Equation 3.10 and Equation 3.12, respectively, the $q(x)$, β and $\pi(x)$ parameters of a excessive-zero Discrete Weibull linear regression model can be related to the set of covariates as follows

$$\begin{aligned} \log\left(-\log\left(q(x)\right)\right) &= \theta_0 + \theta_1 \text{age} + \theta_2 \text{female} + \\ &\quad \theta_3 \text{hhinc} + \theta_4 \text{educ} + \theta_5 \text{work} \\ \log(\beta) &= \vartheta_0 \\ \text{logit}\left(\pi(x)\right) &= \gamma_0 + \gamma_1 \text{age} + \gamma_2 \text{female} + \\ &\quad \gamma_3 \text{hhinc} + \gamma_4 \text{educ} + \gamma_5 \text{work} \end{aligned} \tag{3.23}$$

As a comparison we consider the zero inflated and hurdle models via the Poisson and the Negative Binomial distribution implemented via the function `zeroinfl` and `hurdle` respectively, and available in the `R` package `pscl` [55]. The parameters estimates and the AIC values are presented in Table 3.11 for the zero inflated and hurdle models, respectively. The coefficient of the Discrete Weibull model are parametrised with respect to the logarithm of the median. We note that all the models identify as significant the same variables, i.e. the age, the gender and the income of the patient, while the education and the working condition do not affect significantly the number of visits to a doctor.

German socio-economic panel

We compare our analysis to the study of [113] which fit a zero inflated generalised Poisson regression model to investigate the German Socio-economic Panel (GSOEP). This is an unbalanced panel of 7,293 individual families over 7 years. As in [113] we subset the first 438 individuals, and we aim to predict the number of doctor visits in the last three months using as covariates the gender, the age in years, the health satisfaction, the working condition, the marital status, and the years of schooling. The response variable has approximately 45% of zeros, a mean of 2.93 and a variance of 33.1, thus a dispersion of 11.32. The observed frequencies are shown in Figure 3.20.

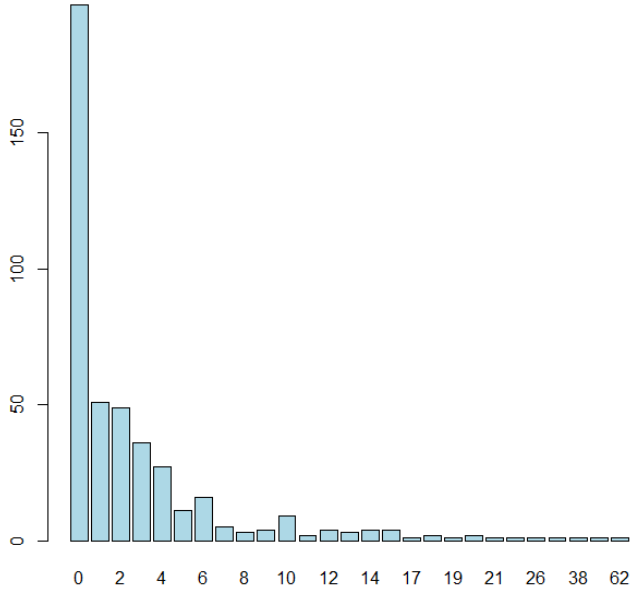


FIGURE 3.20: Bar plot of the number of doctor visit for the GSOEP data.

The $q(x)$, β and $\pi(x)$ parameters of a excessive-zero Discrete Weibull linear regression model can be related to the set of covariates as follows

$$\begin{aligned} \log(-\log(q(x))) &= \theta_0 + \theta_1 \text{female} + \theta_2 \text{age} + \theta_3 \text{hsat} + \\ &\quad \theta_4 \text{married} + \theta_5 \text{working} + \theta_6 \text{educ} \\ \log(\beta) &= \vartheta_0 \\ \text{logit}(\pi(x)) &= \gamma_0 + \gamma_1 \text{female} + \gamma_2 \text{age} + \gamma_3 \text{hsat} + \\ &\quad \gamma_4 \text{married} + \gamma_5 \text{working} + \gamma_6 \text{educ} \end{aligned} \tag{3.24}$$

The parameter estimates and the AIC values via the zero inflated and hurdle Poisson, Negative Binomial and Discrete Weibull regression model are presented in Table 3.12 together with the model significant covariates identified by fitting a zero inflated generalised Poisson model as presented in Table 7 of [113]. The estimates for the Discrete Weibull models are parametrised with respect to the logarithm of the median. We note that the zero inflated and hurdle models via a Poisson, Negative Binomial and Discrete Weibull models detect as significant the same parameters identified by the zero inflated Generalised-Poisson of [113], i.e. the gender and the health satisfaction. Considering the log-likelihood and the AIC values, both the zero inflated and hurdle formulations via the Discrete Weibull outperform the other models.

TABLE 3.12: Parameter estimates and AIC for the count part of the zero-inflated and hurdle model on the GSOEP data. The first column reports the significant variables of the zero-inflated generalised Poisson model taken from Table 7 of [113].

	ZI GPO	ZI PO	ZI NB	ZI DW	hurdle PO	hurdle NB	hurdle DW
(Intercept)	2.53***	2.741***	2.283***	1.926***	2.708***	2.259***	1.824**
	-	(0.237)	(0.661)	(0.62)	(0.236)	(0.672)	(0.635)
female	0.59***	0.276***	0.37*	0.345*	0.274***	0.32.	0.285.
	-	(0.062)	(0.161)	(0.131)	(0.063)	(0.163)	(0.137)
age	-	-0.004	-0.001	-0.001	-0.003	0	0.001
	-	(0.003)	(0.007)	(0.006)	(0.003)	(0.007)	(0.006)
hsat	-0.26***	-0.221***	-0.24***	-0.213***	-0.221***	-0.251***	-0.225***
	-	(0.012)	(0.033)	(0.03)	(0.012)	(0.035)	(0.031)
married	-	0.25***	0.218	0.19	0.251***	0.195	0.159
	-	(0.062)	(0.163)	(0.14)	(0.062)	(0.166)	(0.139)
working	-	0.143*	0.269	0.204	0.15*	0.343.	0.292.
	-	(0.066)	(0.18)	(0.158)	(0.067)	(0.185)	(0.153)
educ	-	-0.005	0.003	0.002	-0.004	0.003	0.002
	-	(0.012)	(0.03)	(0.026)	(0.012)	(0.03)	(0.026)
other	-	-	σ=-0.078	β=0.933***	-	σ=-0.133	β=0.908***
	-	-	(0.203)	(0.081)	-	(0.224)	(0.083)
AIC	1755.01	2254.91	1750.61	1750.85	2254.642	1748.506	1748.424
logLik	-871.50	-1113.46	-860.31	-860.42	-1113.321	-859.253	-859.2121

The coefficients and standard errors (in brackets) are reported.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Unwanted pursuit behaviour In this analysis we follow [62] and investigate the impact of the education level and the level of anxious attachment on the number of unwanted pursuit behaviour perpetration in the context of couple separation trajectories. The response is regressed against the factor education, and the anxious attachment levels. The dataset contains 387 observations with 63.6% of zeros, while the response has a mean of 2.28 and a variance of 23.3, thus a dispersion index of 10.2. The observed frequencies are shown in Figure 3.21.

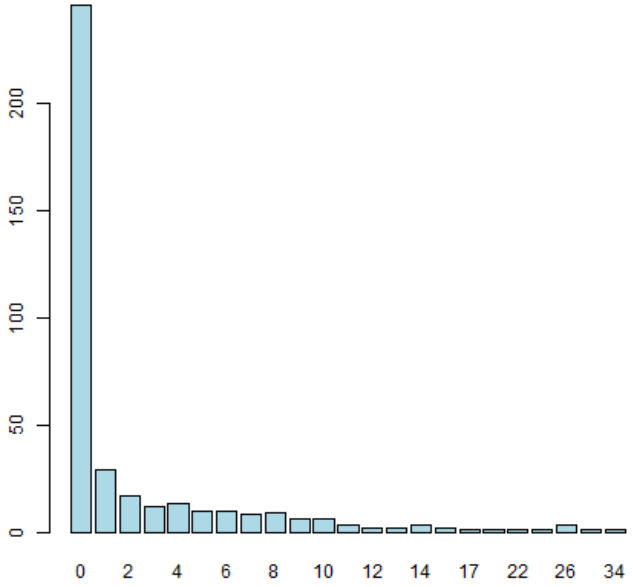


FIGURE 3.21: Bar plot of the number of unwanted pursuit behaviour perpetration in the context of couple separation.

To model these data [88] employed a zero-inflated COM–Poisson model. Here we consider an excessive-zero Discrete Weibull linear regression model where $q(x)$, β and $\pi(x)$ parameters can be related to the set of covariates as follows

$$\begin{aligned} \log\left(-\log\left(q(x)\right)\right) &= \theta_0 + \theta_1 \text{education} + \theta_2 \text{anxiety} \\ \log(\beta) &= \vartheta_0 \\ \text{logit}\left(\pi(x)\right) &= \gamma_0 + \gamma_1 \text{education} + \gamma_2 \text{anxiety} \end{aligned} \tag{3.25}$$

We report the results of [88] in Table 3.13, together with the Discrete Weibull model estimates parametrised to the logarithm of the median. The results show a good performance of the zero inflated Discrete Weibull and zero inflated Geometric model. Given that the Geometric distribution can be seen as a special case of the Discrete Weibull as detailed in section 3.1, this result confirms the potential of the Discrete Weibull distribution in modelling count data. Moreover, all the zero-inflated models identify the education as significant factor in predicting the behaviour of individuals in the context of couple separation.

TABLE 3.13: Parameter estimates and AIC for the count part of the zero-inflated and hurdle model on the number of unwanted pursuit behaviour perpetration. The results for the models other than Discrete Weibull are taken from Table 2 of [88].

	ZI PO	ZI NB	ZI CMP	ZI G	ZI DW	hurdle PO	hurdle NB	hurdle DW
(Intercept)	1.921*	1.723*	-0.160*	1.770*	1.365***	1.921***	1.725***	1.368***
	(0.044)	(0.15)	(0.077)	(0.122)	(0.313)	(0.044)	(0.148)	(0.314)
Education	-0.350*	-0.490*	-0.068*	-0.476*	-0.454*	-0.35***	-0.487*	-0.45*
	(0.071)	(0.206)	(0.034)	(0.191)	(0.178)	(0.071)	(0.206)	(0.178)
Anxiety	0.133*	0.205	0.023	0.199	0.206*	0.133***	0.207.	0.207*
	(0.034)	(0.108)	(0.015)	(0.1)	(0.092)	(0.034)	(0.107)	(0.091)
other	-	σ=0.821	σ=0.001	-	β=0.915***	-	σ=-0.187	β=0.918***
	-	(0.226)	(0.031)	(0.104)	-	-	(0.273)	(0.105)
AIC	1616.9	1266.3	1268.3	1264.8	1265.9	1616.921	1266.526	1266.2
logLik	-802.45	-626.14	-627.17	-626.42	-625.98	-802.461	-626.263	-626.104

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of fish caught The data is available at "https://stats.idre.ucla.edu/stat/data/fish.csv". The study focus on 250 groups that went to a park, and each group was questioned about how many fish they caught, how many children, and how many people were in the group, and whether or not they brought a camper to the park. In addition to predicting the number of fish caught, there is interest in predicting the existence of excess zeros, i.e. the probability that a group caught zero fish. The dataset contains 56.8% of zeros, while the response has a mean of 3.29 and a variance of 135.37, thus a dispersion index of approximately 41. The observed frequencies are shown in Figure 3.22.

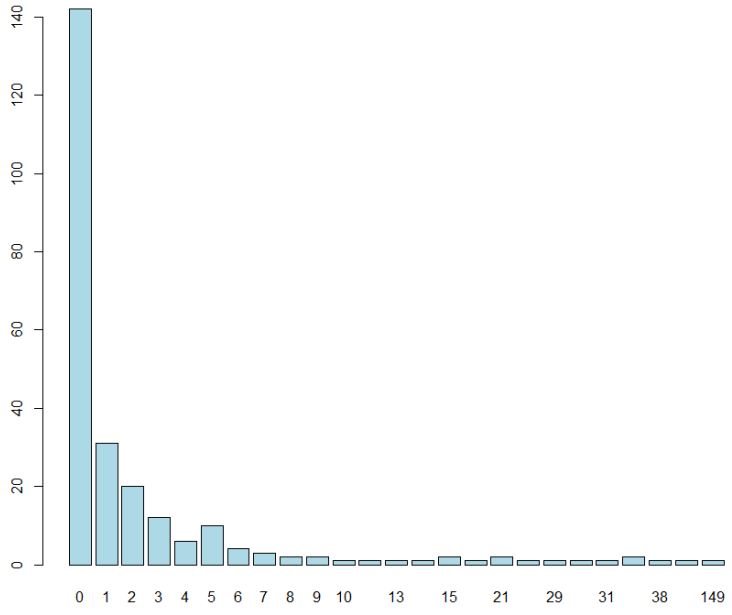


FIGURE 3.22: Bar plot of the number of fish caught data.

Thus, the $q(x)$, β and $\pi(x)$ parameters of a excessive-zero Discrete Weibull linear regression model can be related to the set of covariates as follows

$$\begin{aligned} \log(-\log(q(x))) &= \theta_0 + \theta_1 \text{child} + \theta_2 \text{camper1} + \theta_3 \text{persons} \\ \log(\beta) &= \vartheta_0 \\ \text{logit}(\pi(x)) &= \gamma_0 + \gamma_1 \text{child} + \gamma_2 \text{camper1} + \gamma_3 \text{persons} \end{aligned} \tag{3.26}$$

The parameter estimates parametrised to the logarithm of the median and the AIC values via the zero inflated and hurdle models Discrete Weibull models are presented in Table 3.14. Once again, in terms of AIC the zero inflated and the hurdle Discrete Weibull models outperforms all the alternatives. The zero inflated COM-Poisson regression model has been computed via the `R` function `glm.cmp` available in the `R` package `COMPoissonReg`. We note that the COM-Poisson model shows better performance than the Poisson model.

TABLE 3.14: Parameter estimates and AIC for the count part of the zero-inflated and hurdle model on the number of fish caught data.

	ZI PO	ZI CMP	ZI NB	ZI DW	hurdle PO	hurdle NB	hurdle DW
(Intercept)	-0.798***	-0.678	-1.618***	-1.798***	-0.826***	-1.622**	-2.456***
	(0.171)	(0.085)	(0.32)	(0.194)	(0.172)	(0.596)	(0.351)
child	-1.137***	-0.139	-1.261***	-1.234***	-1.139***	-1.095***	-1.156***
	(0.093)	(0.047)	(0.247)	(0.172)	(0.093)	(0.32)	(0.204)
camper1	0.724***	0.075	0.386	0.404.	0.734***	0.375	0.515
	(0.093)	(0.04)	(0.246)	(0.157)	(0.093)	(0.336)	(0.21)
persons	0.829***	0.145	1.09***	0.958***	0.835***	1.003***	0.983***
	(0.044)	(0.028)	(0.112)	(0.085)	(0.044)	(0.155)	(0.105)
other	-	σ=-0.957	σ=-0.593***	β=0.74***	-	σ=-1.053*	β=0.62***
	-	(0.35)	(0.158)	(0.054)	-	(0.497)	(0.105)
AIC	1521.463	815.937	809.079	802.349	1519.236	808.318	803.942
logLik	-752.732	-398.968	-395.539	-391.998	-751.618	-395.159	-392.971

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.7 Conclusions

The regression framework presented in this chapter via a Discrete Weibull distribution has been successfully applied to a number of real data examples, thus it can be considered a highly competitive alternative to the current models for count data in case of over- or under-dispersion, and in the presence of an excess number of zeros. Such data structures appear frequently in various applications, such as healthcare, social science, psychology, engineering, business, and so on. In addition to the linear regression framework, the model has been extended to consider cases when data are grouped into clusters, or panels, or correlated groups, i.e. hierarchical structures.

Chapter 4

Non-linear models for counts via a Discrete Weibull distribution

In the parametric literature, the inclusion of dependencies other than linear had been addressed by developing new distributions with additional parameters, e.g. the generalised Gamma approach of [74] for continuous responses, or by adopting more flexible non-linear regression models that can link all parameters of the distribution to the covariates, most notably the generalized additive models for location, scale and shape i.e. GAMLSS of [84].

In chapter 3 we employed GLMs where the conditional distribution of the response variable given the predictors is assumed to follow a specified distribution with the conditional mean linked to the predictors via a regression model. Specifically, we employed the Negative Binomial and the Poisson-inverse Gaussian models in the case of over-dispersed data, the generalised Poisson in the case of under-dispersed data, the Poisson and COM-Poisson models in both the cases, and the zero inflated and hurdle specifications of these models when in the presence of excessive zeros. Thus, we compared these models with our approach via a Discrete Weibull distribution which considered a linear relationship between the linked-transformed parameters and the covariates. So far, we have kept the second distributional parameter of the Discrete Weibull fixed. Here we extend our parametric approach by including non-linear dependencies for both the regression parameters and the covariates. In this way we are able to model more accurately the full conditional distribution of Y given X, i.e. all conditional quantiles.

4.1 The GAMLSS Discrete Weibull regression model

By assuming that the response variable has a discrete Weibull distribution conditional on the exogenous variables, we employ generalised additive models to link the parameters to the predictors. In this sense, this approach could offer an alternative to the more traditional non-parametric quantile regression models, which are rather challenging for counts. Moreover, adding a link to both parameters means that conditional quantiles of various shapes and complexity can be captured. Specifically, we assume that the response $Y|X$ has a Discrete Weibull conditional distribution, with the q and

β parameters linked to the covariates x via a generalized additive model as follows

$$\begin{aligned}\log\left(-\log\left(q(x)\right)\right) &= \sum_{p=1}^{P}\sum_{d=0}^{D_p}\theta_{0pd}{x_p}^d + \sum_{p=1}^{P}\sum_{k=1}^{K_p}\theta_{pk}(x_p - h_{pk})^{D_p} I(x_p > h_{pk}),\\ \log\left(\beta(x)\right) &= \sum_{p=1}^{P}\sum_{d=0}^{D'_p}\vartheta_{0pd}{x_p}^d + \sum_{p=1}^{P}\sum_{k=1}^{K'_p}\vartheta_{pk}(x_p - h_{pk})^{D'_p} I(x_p > h_{pk}),\end{aligned} \tag{4.1}$$

where $x = (1, x_1, \ldots, x_P)$, D and D' denote the degrees of the polynomials, K and K' the number of break points or internal knots, $I(\cdot)$ is the indicator function, so $I(x_p > h_{pk})$ is 1 if $x_p > h_{pk}$ and 0 otherwise, and (θ, ϑ) are the vectors of parameters to be estimated. The general formulation as a B-spline model [29] includes models of varying complexity, such as linear models with or without interactions and orthogonal polynomial basis [99]. This in turn returns conditional quantiles of various shapes and complexity. Rather than defining the number of knots and degrees, is also possible to formulate the problem as a penalized regression spline [112].
We can now extend the formulation of the τ-quantile function presented in Equation 3.5 linking both the parameter q and β to the covariates x. Thus, for a fixed quantile $\tau \in [0,1]$, the $\mu^{(\tau)}$ quantile of a $Y|X \sim$ Discrete Weibull is given by

$$\mu^{(\tau)} = \lceil \mu^{*(\tau)} \rceil = \left\lceil \left(\frac{\log(1-\tau)}{\log(q(x))} \right)^{\frac{1}{\beta(x)}} - 1 \right\rceil. \tag{4.2}$$

From this,

$$\log(\mu^{*(\tau)} + 1) = \frac{1}{\beta(x)}\log(-\log(1-\tau)) - \frac{1}{\beta(x)}\log(-\log(q(x))). \tag{4.3}$$

The formulation of the log-quantile given in Equation 4.3 will be used to graphically inspect the quantiles of the models.

Considering one covariate x only, and dropping the indices p of the model for simplicity, we look closely at three cases to inspect the level of flexibility of a Discrete Weibull model in approximating conditional distributions.

- **Discrete Weibull linear regression model with β constant.**
  This model is specified as in Equation 4.1 with $D = 1$, $D' = 0$ and no knots, i.e.:

  $$\begin{aligned}\log\left(-\log\left(q(x)\right)\right) &= \theta_{00} + \theta_{01}x\\ \log\left(\beta\right) &= \vartheta_{00}.\end{aligned} \tag{4.4}$$

  The top-left plot in Figure 4.1 shows the case $\theta_{00} = -10$, $\theta_{01} = -5$, $\vartheta_{00} = 0.7$. The figure plots $\log(\mu^{(*\tau)} + 1)$ from Equation 4.3. As expected by Equation 4.3, a linear model with β constant returns log-quantiles which are linear and parallel. This is the case of the models considered in chapter 3.

- **Discrete Weibull linear regression model with β not constant.**
  This model is specified as in Equation 4.1 with $D = D' = 1$ and no knots, for example:

  $$\begin{aligned}\log\left(-\log\left(q(x)\right)\right) &= \theta_{00} + \theta_{01}x\\ \log\left(\beta(x)\right) &= \vartheta_{00} + \vartheta_{01}x,\end{aligned} \tag{4.5}$$

  for the case of a linear model on both $q(x)$ and $\beta(x)$. The top-right plot in Figure 4.1 shows the case $\theta_{00} = -5$, $\theta_{01} = -10$, $\vartheta_{00} = -1.5$, $\vartheta_{01} = 3$. This plot shows how a non-constant $\beta(x)$ allows to obtain log-quantiles that are not parallel.

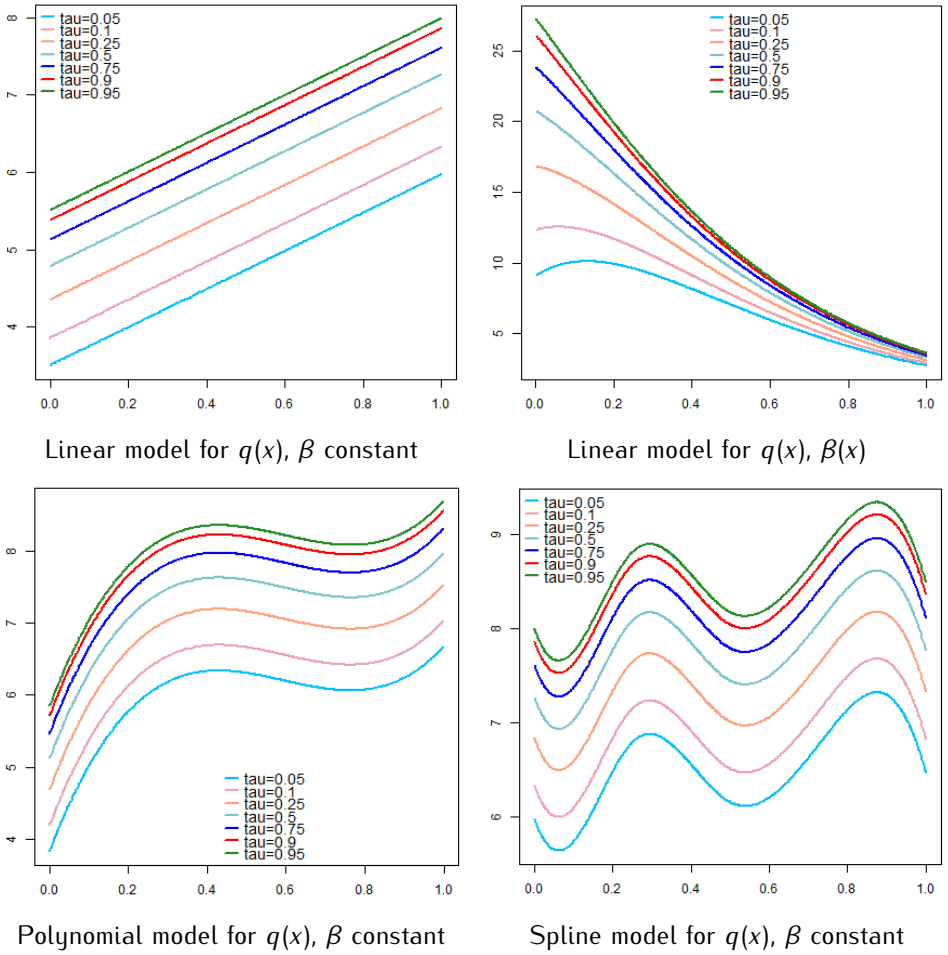


FIGURE 4.1: Plot of the conditional quantiles for Discrete Weibull models under linear (top) and non-linear (bottom) models, and β fixed (top left, bottom) and not (top right).

- **Discrete Weibull non-linear regression model.**

  Here there are two cases of interest: a parametric polynomial model and a spline model. For example, setting $D = 3$, $D' = 0$ and no knots in Equation 4.1 leads to a 3rd-degree orthogonal polynomial model for $q(x)$:

$$\begin{aligned} \log(-\log(q(x))) &= \theta_{00} + \theta_{01}x + \theta_{02}x^2 + \theta_{03}x^3 \\ \log(\beta) &= \vartheta_{00}. \end{aligned} \tag{4.6}$$

  The bottom-left plot in Figure 4.1 shows the quantiles for the 3rd-degree polynomial model with $\theta_{00} = -15$, $\theta_{01} = -25$, $\theta_{02} = 20$, $\theta_{03} = -18$, and $\vartheta_{00} = 0.7$. On the other hand, setting $D = K = 3$, $D' = 0$ in Equation 4.1 leads to a B-spline model for $q(x)$ with three interior knots:

$$\begin{aligned} \log(-\log(q(x))) = \theta_{00} + \theta_{01}x + \theta_{02}x^2 + \theta_{03}x^3 + \theta_1(x-h_1)^3 I(x > h_1) + \\ + \theta_2(x-h_2)^3 I(x > h_2) + \theta_3(x-h_3)^3 I(x > h_3) \\ \log(\beta) = \vartheta_{00}. \end{aligned} \tag{4.7}$$

  Cubic splines are typically complex enough for most real applications [31]. The degrees of freedom for the fitted model are given by S=1+D+K. The knots are typically evenly spaced throughout the range of observed values or placed at some quantiles of the variable of interest. To generate the smooth term of x to pass into the model formula, in `R` software we employ the `bs` function available in the `splines` package. This generates the the B-spline basis

matrix of the piecewise polynomial term x with the specified number of interior knots K and degree D. The bottom-right plot in Figure 4.1 shows the quantiles for the cubic spline model having set $\theta_{00} = -15$, $\theta_{01} = -13$, $\theta_{02} = -19$, $\theta_{03} = -14$, $\theta_1 = -17$, $\theta_2 = -18.5$, $\theta_3 = -16$, and $\vartheta_{00} = 0.7$. The cubic spline, together with the assumption of a constant β, leads to parallel and non-linear log-quantiles, as expected by Equation 4.3.

4.2 Parameter estimation

Parameter estimation is done via maximum likelihood, as presented in section 3.3. In addition, we consider two main extensions. For high dimensional problems or when variable selection is of interest we consider the L_1 penalty as in the least absolute shrinkage and selection operator widely known as Lasso ([101], or see Chapter 3 in [40]). By retaining a subset of the predictors and discarding the rest, this method is very efficient also for large problems, e.g. when the number of variables is larger than the number of observations, when the usual maximum likelihood approach will fail. The second extension considers a local approach via a weighting function or kernel which assign a weight to x_i based on its distance from a specified point x^0 (see chapter 6 in [40]). We described both methods in the next two sections.

4.2.1 The L_1 penalised Discrete Weibull regression model

Lasso is a regression method which involves penalizing the absolute size of the regression coefficients. The penalisation will results in some of the parameter being exactly zero. This is convenient when we want some automatic variable selection, or when dealing with highly correlated predictors. For the above reasons, we extended the regression model in Equation 4.1 with the inclusion of a L_1 penalty term for the selection of the variables. The L_1 penalized estimation method shrinks the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. In particular, the parameters of the model presented in Equation 4.1 are now estimated by maximising the weighted log-likelihood with L_1 penalty

$$\sum_{i=1}^{n} l(y, x; \theta, \vartheta) + \lambda_1 \sum_{p=1}^{P} \sum_{d=0}^{D_P} |\theta_{0pd}| + \lambda_2 \sum_{p=1}^{P} \sum_{k=1}^{K_P} |\theta_{pk}| + \lambda_3 \sum_{p=1}^{P} \sum_{d=0}^{D'_P} |\vartheta_{0pd}| + \lambda_4 \sum_{p=1}^{P} \sum_{k=1}^{K'_P} |\vartheta_{pk}|.$$

We note that as the λs terms increase, more coefficients are set to zero, i.e. less variables are selected, and among the non-zero coefficients, more shrinkage is employed. In `R` software and within the `gamlss` regression function, the estimation of the penalised smoothing coefficients for each distributional parameter with respect to the xth term can be done via the `ri` function with `Lp=1` penalty.

4.2.2 The Discrete Weibull regression model with Gaussian kernel weights

For the local estimator we extended the Discrete Weibull likelihood with the inclusion of Gaussian kernel weights. The kernel smoothing is one of the most widely used non-parametric data smoothing techniques. The functional form of the kernel implies that the weights are much larger for the observations where x_i is close to a pre-specified x^0. The size of the weights is parametrized by the bandwidth b, where a very large bandwidth leads to a very smooth model. For each chosen bandwidth b, the parameters $\theta^{(b)}$ and $\vartheta^{(b)}$ of Equation 4.1 are estimated by optimising

$$\sum_{i=1}^{n} w_i(x^0, b) l(y, x|\theta, \vartheta),$$

with the weights w_i local around the vector of predictors x^0 and dependent on a bandwidth b via the Gaussian kernel

$$w_i(x^0, b) = \exp\left(-\frac{\|x_i - x^0\|}{2b^2}\right),$$

where

$$\|x_i - x^0\| = \sum_{p=1}^{P}\left(\frac{x_{ip} - x_p^0}{\mathrm{sd}(x_p)}\right)^2.$$

For the selection of the bandwidth of a Gaussian kernel density estimator one can employ the `R` function `bw.nrd0` available in the `stats` package to select the optimally smoothed curve i.e. the density estimate which is close to the true density, thus avoiding curves which are under-smoothed since contains too many spurious data, or curves which are over-smoothed which will fail in the identification of the underlying structure of the data.

4.3 Model selection and comparison

We compare our Discrete Weibull specification with existing parametric approaches which employ different distributions and with the jittering approach of [63]. We measure the performance of these approaches by considering three different quantiles, namely $\tau = 0.25, 0.5, 0.75$. Thus, in subsection 4.5.1 for each τ and for each model we will evaluate the accuracy in the estimation of the conditional quantile by calculating the root mean squared error

$$\mathrm{RMSE} = \left(\frac{\sum_{i=1}^{n}\left(\hat{\mu}_i^{(\tau)} - \mu_i^{(\tau)}\right)^2}{n}\right)^{0.5}, \tag{4.8}$$

where $\mu_i^{(\tau)}$ is the real quantile, and $\hat{\mu}_i^{(\tau)}$ is the fitted quantile from the specified model. For the Discrete Weibull model, $\hat{\mu}_i^{(\tau)}$ and $\mu_i^{(\tau)}$ are calculated as in Equation 4.3 using the model fitted values from Equation 4.1, i.e. $\hat{q}(x)$ and $\hat{\beta}(x)$, and the real values, i.e. $q(x)$ and $\beta(x)$, from the true parameters, respectively. For the other models, we use the functions `qNBI`, `qGPO`, `qPIG` and `qPO` in the `R` package `gamlss.dist` and the function `qcmp` available in the `R` package `COMPoissonReg` to calculate the quantiles of the Negative Binomial, generalised Poisson, Poisson-inverse Gaussian, Poisson and COM-Poisson model, respectively.

In real data applications, as described in section 3.4 we employ the AIC estimator for the model selection. Moreover, we consider the partial effects in order to quantify the change in the quantiles of the dependent variable in response to a change in each explanatory variable, while keeping all the other covariates constant. In particular, let x^0 denotes the vector of predictors, where each predictor is set to their sample mean $\bar{x}$ if continuous and to their mode if dummy. Then, the effect for the regressor x_p is calculated as the difference $\mu^{*(\tau)}(x_p^1) - \mu^{*(\tau)}(x^0)$, where $\mu^{*(\tau)}(x_p^1)$ is the quantile estimated on the vector x_p^1, which is equal to x^0, with the exception of the p^{th} variable which is increased by one unit, while $\mu^{*(\tau)}(x^0)$ is the fitted quantile on x^0.

4.4 Model diagnostic

In addition to the diagnostic plot based on the randomised quantiles residuals as detailed in section 3.4, in section 4.6 we also assess the goodness-of-fit of the model following the approach of [74]. In particular, one would expect $100\left(\hat{\mu}_{i+1}^{(\tau)} - \hat{\mu}_i^{(\tau)}\right)$ % of the data to lie between the i^{th} and the $(i+1)^{th}$ conditional quantile, so we compare this target

value with that obtained using the estimated conditional quantiles. Although this approach requires continuous response data, it works well on the examples reported in this analysis where the response variable takes an enough large number of discrete values. We consider the ten regions defined by the 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% quantiles.

4.5 Simulation study

We perform a number of simulations to compare the Discrete Weibull generalised additive regression with parametric alternatives and with the Jittering approach of [64]. Specifically, the parametric comparison is done in `R` by fitting the Poisson, Negative Binomial, generalised Poisson, and Poisson-Inverse Gaussian via the `gamlss` function, while the COM-Poisson regression is implemented via the `R` function `glm.cmp` within the `COMPoissonReg` package. The Jittering approach can be fit in `R` via the `lqm.counts` and `rq.counts` functions in the `lqmm` and `Qtools` packages, respectively. The algorithm implemented in the `lqm.counts` function is based on a Laplace gradient estimation, whereas the one implemented in the `rq.counts` function is based on a linear programming estimation. In this chapter, we use the `rq.counts` function, which provides more stable estimates for small sample sizes. Moreover, we employ generalised additive regression models extended to all the distributional parameters for the distributions presented above. Across the simulations and the different models, we use generalized additive models of the same complexity for a fair comparison.

4.5.1 Simulated data from a Discrete Weibull model

We first simulate data from our proposed model as in Equation 4.3. In particular, we set n=50,100,1000, $x \sim$ Uniform$(0,1)$, $Y|X \sim$ Discrete Weibull$(q(x),\beta(x))$, and we consider the following cases:

- Case (1). Linear model for $q(x)$, β constant.

$$\begin{aligned} &\log(-\log(q(x))) = -3.5 - x, \\ &\quad \text{(a)} \quad \log(\beta) = 0.7 \quad \rightarrow \quad \text{over-dispersed} \\ &\quad \text{(b)} \quad \log(\beta) = 1.1 \quad \rightarrow \quad \text{under-dispersed.} \end{aligned} \tag{4.9}$$

- Case (2). Linear model for $q(x)$, $\beta(x)$.

$$\begin{aligned} &\log(-\log(q(x))) = -5 - 3x \\ &\text{(a)} \quad \log(\beta(x)) = 0.9 + 0.3x \quad \rightarrow \quad \text{over-dispersed} \\ &\text{(b)} \quad \log(\beta(x)) = 1.2 + 0.5x \quad \rightarrow \quad \text{under-dispersed.} \end{aligned} \tag{4.10}$$

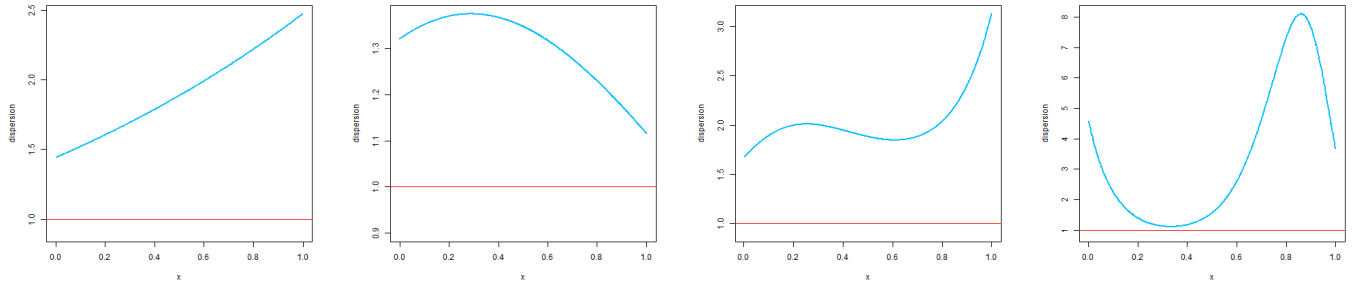
- Case (3). Third degree polynomial model for $q(x)$, β constant.

$$\begin{aligned} &\log(-\log(q(x))) = -5 - 7x - 4x^2 - 6x^3, \\ &\quad \text{(a)} \quad \log(\beta) = 0.8 \quad \rightarrow \quad \text{over-dispersed} \\ &\quad \text{(b)} \quad \log(\beta) = 1.6 \quad \rightarrow \quad \text{under-dispersed.} \end{aligned} \tag{4.11}$$

- Case (4). Cubic spline model for $q(x)$, β constant.

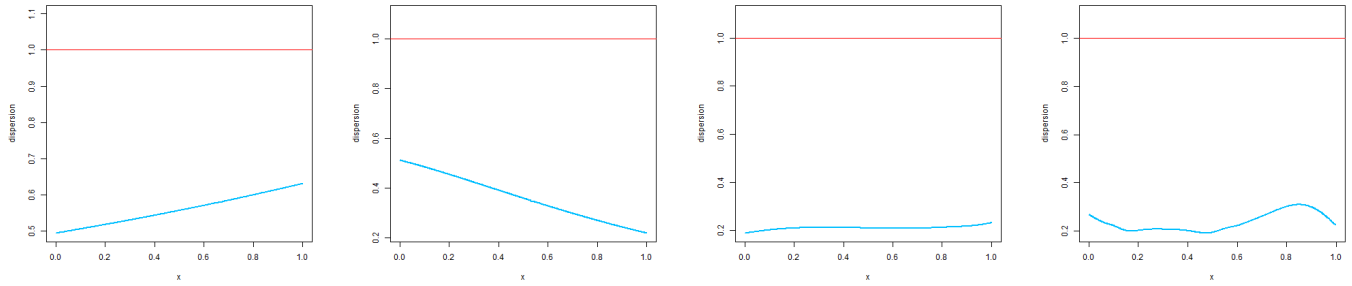
$$\begin{aligned} \log(-\log(q(x))) &= -7 - 5x - 3x^2 - 4x^3 - 8(x-h_1)^3 I(x > h_1) + \\ &\quad - 9(x-h_2)^3 I(x > h_2) - 6(x-h_3)^3 I(x > h_3), \\ \text{(a)} \quad \log(\beta) &= 0.8 \quad \rightarrow \quad \text{over-dispersed} \\ \text{(b)} \quad \log(\beta) &= 1.6 \quad \rightarrow \quad \text{under-dispersed.} \end{aligned} \tag{4.12}$$

Setting the values as above leads to a range of values of $q(x)$ between 0.79 and 1 and a range of values of $\beta(x)$ between 2 and 8.2, thus allowing us to explore the fit of the models for a number of different quantiles. The four cases correspond to models of varying complexity. In addition, we also consider both cases of over-dispersion and of under-dispersion relative to Poisson. Figure 4.2 shows the conditional dispersions for the over-dispersed cases, whereas Figure 4.3 is for the under-dispersed cases. The red line in the graphs represents the threshold value of dispersion 1. The plots show cases of either over- or under-dispersion for all values x. In fact, a Discrete Weibull regression model can capture also cases of mixed dispersion, with over-dispersion for some covariates' patterns and under-dispersion for others.



Linear model for $q(x)$, β constant | Linear model for $q(x)$ and $\beta(x)$ | Polynomial model for $q(x)$, β constant | Spline model for $q(x)$, β constant

FIGURE 4.2: Plot of the conditional dispersion values for the cases of over-dispersion of Discrete Weibull simulated data.



Linear model for $q(x)$, β constant | Linear model for $q(x)$ and $\beta(x)$ | Polynomial model for $q(x)$, β constant | Spline model for $q(x)$, β constant

FIGURE 4.3: Plot of the conditional dispersion values for the cases of under-dispersion of Discrete Weibull simulated data.

By fitting the models in Equation 4.9, Equation 4.10, Equation 4.11, and Equation 4.12, for the over- and under-dispersed data respectively, we note the computational gain in terms of CPU time in seconds by using our parametric approach via the Discrete Weibull model with respect to the COM-Poisson model and the Jittering method for the 50% τ-quantile averaged over 50 dithered samples, as presented in Table 4.1.

TABLE 4.1: System time (in seconds) performance comparison between the same specification model via different approaches on over- and under-dispersed data simulated from a Discrete Weibull model under four different specifications: (1) linear link on q, constant β, (2) linear link on both q and β, (3) cubic polynomial link on q, constant β, (4) cubic spline on q, constant β

over-dispersed				
	Case (1)	Case (2)	Case (3)	Case (4)
DW	0.16	0.16	0.14	0.18
NB	0.03	0.05	0.03	0.06
PIG	0.07	0.06	0.06	0.14
CMP	2.64	5.7	9.59	18.35
GPO	0.09	0.09	0.08	0.11
PO	0.01	0.02	0.02	0.01
Jittering	0.42	0.39	0.42	0.47
under-dispersed				
	Case (1)	Case (2)	Case (3)	Case (4)
DW	0.21	0.19	0.19	0.24
CMP	2.26	6.85	7.08	19.23
GPO	0.11	0.12	0.11	0.14
PO	0.02	0.01	0.01	0.01
Jittering	0.4	0.41	0.4	0.47

Table 4.2 and Table 4.3 report the errors, calculated as in Equation 4.8, averaged over 100 iterations, for the three different quantiles $\tau = (0.25, 0.5, 0.75)$ and for sample sizes $n = (50, 100, 1000)$, for the over-dispersed and under-dispersed cases, respectively.

TABLE 4.2: Comparison of different models in terms of root mean squared error on over-dispersed data simulated from a Discrete Weibull model under four different model specifications: (1) linear link on q, constant β, (2) linear link on both q and β, (3) cubic polynomial link on q, constant β, (4) cubic spline on q, constant β.

	Jittering			DW			PO			PIG			CMP			NB		
τ\n	50	100	1000	50	100	1000	50	100	1000	50	100	1000	50	100	1000	50	100	1000
(1)																		
0.25	0.782	0.578	0.354	0.652	0.520	0.300	1.123	1.022	0.995	0.944	0.82	0.847	0.961	0.872	0.970	0.698	0.556	0.386
0.500	0.934	0.678	0.399	0.757	0.605	0.336	0.787	0.642	0.496	1.347	1.265	1.359	1.385	1.293	1.359	0.730	0.604	0.364
0.750	1.116	0.820	0.429	0.958	0.708	0.394	0.966	0.883	0.743	1.834	1.676	1.783	1.760	1.670	1.725	0.947	0.704	0.412
(2)																		
0.250	0.748	0.651	0.400	0.697	0.581	0.363	0.755	0.614	0.549	1.216	1.255	1.276	1.321	1.293	1.313	0.691	0.569	0.365
0.500	0.879	0.729	0.464	0.742	0.635	0.449	0.708	0.626	0.466	1.316	1.345	1.390	1.394	1.377	1.345	0.710	0.627	0.507
0.750	0.843	0.753	0.490	0.789	0.668	0.447	0.822	0.771	0.661	1.46	1.502	1.457	1.532	1.575	1.507	0.806	0.709	0.520
(3)																		
0.250	1.747	1.596	0.531	1.158	0.948	0.416	1.938	2.481	1.125	6.173	5.387	1.297	6.375	5.727	1.376	1.268	1.230	0.466
0.500	1.979	1.826	0.482	1.590	1.216	0.432	1.778	1.505	0.537	8.911	7.726	1.779	8.953	7.873	1.824	1.637	1.336	0.428
0.750	2.594	1.868	0.560	2.158	1.555	0.508	3.512	2.138	0.966	12.01	10.37	2.382	11.79	10.21	2.349	2.470	1.664	0.554
(4)																		
0.250	3.768	2.463	0.765	2.187	1.543	0.475	4.630	4.280	3.128	9.310	9.790	7.220	9.880	10.49	7.673	2.591	1.738	0.639
0.500	3.841	2.707	0.808	2.823	2.056	0.575	3.133	2.244	0.780	13.12	13.73	10.38	13.47	14.19	10.71	2.989	2.078	0.715
0.75	4.534	3.505	0.906	3.660	2.763	0.735	4.407	3.964	2.630	17.62	18.50	14.05	17.46	18.32	13.95	3.793	2.846	0.971

TABLE 4.3: Comparison of different models in terms of root mean squared error on under-dispersed data simulated from a Discrete Weibull model under four different model specifications: (1) linear link on q, constant β, (2) linear link on both q and β, (3) cubic polynomial link on q, constant β, (4) cubic spline on q, constant β.

	Jittering			DW			PO			CMP			GPO		
τ\n	50	100	1000	50	100	1000	50	100	1000	50	100	1000	50	100	1000
(1)															
0.25	0.313	0.230	0.058	0.257	0.208	0.052	0.511	0.492	0.570	0.220	0.150	0.020	0.290	0.340	0.330
0.5	0.453	0.409	0.249	0.406	0.381	0.214	0.456	0.376	0.376	0.542	0.537	0.535	0.533	0.575	0.561
0.75	0.468	0.396	0.233	0.424	0.359	0.197	0.499	0.498	0.494	0.667	0.639	0.673	0.690	0.671	0.687
(2)															
0.25	0.546	0.443	0.318	0.440	0.369	0.216	0.801	0.775	0.719	1.298	1.290	1.341	1.365	1.430	1.462
0.5	0.550	0.495	0.280	0.496	0.433	0.234	0.566	0.523	0.457	1.459	1.459	1.497	1.490	1.484	1.537
0.75	0.497	0.374	0.215	0.474	0.362	0.179	0.650	0.609	0.616	1.530	1.473	1.537	1.639	1.549	1.680
(3)															
0.25	0.138	0.139	0.008	0.136	0.121	0.002	0.635	0.584	0.809	0.120	0.090	0.060	1.124	0.792	0.812
0.5	0.414	0.289	0.026	0.306	0.287	0.004	0.406	0.265	0.176	0.360	0.260	0.340	1.129	0.652	0.182
0.75	0.537	0.478	0.434	0.480	0.425	0.440	0.675	0.664	0.767	0.591	0.576	0.602	1.296	0.890	0.779
(4)															
0.25	0.546	0.443	0.318	0.440	0.369	0.216	0.801	0.775	0.719	1.298	1.290	1.341	1.365	1.430	1.462
0.5	0.550	0.495	0.280	0.496	0.433	0.234	0.566	0.523	0.457	1.459	1.459	1.497	1.490	1.484	1.537
0.75	0.497	0.374	0.215	0.474	0.362	0.179	0.650	0.609	0.616	1.530	1.473	1.537	1.639	1.549	1.680

TABLE 4.4: Case (4) cubic spline on q, constant β: root mean squared error comparison of linear Discrete Weibull model for $q(x)$ and β constant, and linear Jittering model versus the well-specified Discrete Weibull B-spline model for $q(x)$ and β in case of over- and under-dispersed data.

(4)	linear Jittering			linear DW			DW B-spline		
τ\n	50	100	1000	50	100	1000	50	100	1000
Over-disp.									
0.25	5.521	5.464	4.987	6.155	6.132	6.053	2.240	1.606	0.601
0.5	7.778	7.495	6.930	7.764	7.698	7.427	2.926	2.103	0.740
0.75	9.933	9.548	8.906	9.888	9.803	9.281	3.812	2.782	0.940
Under-disp.									
0.25	0.824	0.793	0.726	0.924	0.938	0.946	0.461	0.366	0.229
0.5	0.809	0.800	0.813	0.792	0.783	0.759	0.510	0.426	0.245
0.75	0.917	0.871	0.843	0.940	0.912	0.890	0.467	0.355	0.185

Considering the case of over-dispersed data, for every τ and independently on the sample size, the Discrete Weibull model outperforms all the models. The two main competitors are the Jittering method and Negative Binomial model. The Poisson model performs better than the the Poisson-Inverse Gaussian and the COM-Poisson model, particularly when non-linear dependencies are considered, i.e. fourth case, but none of them seems to be a valuable alternative to the Discrete Weibull model. Regarding the case of under-dispersion, the Discrete Weibull model outperforms the Jittering and all the other models in all the case, although the error measures of the Jittering and the Discrete Weibull model are very close under these two approaches. In the case of under-dispersed data the COM-Poisson model shows better performance than in the case of over-dispersed data. The results obtained via a Poisson and a generalised Poisson are very similar. Table 4.4 focusses only on the fourth case and compares the Discrete Weibull well-specified non-linear model with the simpler (miss-specified) linear Discrete Weibull model for $q(x)$ and β used in chapter 3, and the linear Jittering model. The table shows how the linear Discrete Weibull and the Jittering are equally disadvantaged by the miss-specification of the model, although Jittering shows a slightly better performance.

4.5.2 Simulated data from a Poisson and a Negative Binomial model

In order to advocate the use of this approach for general regression problems with a discrete response variable, we test the robustness of the approach to misspecification in the distribution of the response variable. In particular, we consider the cases of Poisson and Negative Binomial. We set n=50,100,1000, $x \sim \text{Uniform}(0,1)$ and consider the following models

Poisson data: $Y|X \sim \text{Poisson}(\mu(x))$:

- Case (1). Linear model.

$$\log(\mu(x)) = 1 - 1.5x.$$

- Case (3). Third degree polynomial model.

$$\log(\mu(x)) = 1 + 1.5x + 0.5x^2 + x^3.$$

- Case (4). Cubic spline model.

$$\log(\mu(x)) = 1.5 + x + 0.6x^2 + 0.9x^3 + 0.5(x-h_1)^3 I(x > h_1) + \\ + 0.7(x-h_2)^3 I(x > h_2) + 0.8(x-h_3)^3 I(x > h_3).$$

Because the Poisson distribution has only one parameter, we do not evaluate the model presented in Case (2). Table 4.5 reports the errors as in Equation 4.8 for the quantiles $\tau = 0.25, 0.5, 0.75$, averaged over 100 iterations, and for sample sizes $n = 50, 100, 1000$.

TABLE 4.5: Comparison of different models in terms of root mean squared error on simulated Poisson data under four different model specifications: (1) linear link on q, constant β, (3) cubic polynomial link on q, constant β, (4) cubic spline on q, constant β.

	Jittering			DW			PO		
τ\n	50	100	1000	50	100	1000	50	100	1000
(1)									
0.25	0.400	0.275	0.090	0.362	0.255	0.074	0.321	0.221	0.070
0.5	0.600	0.431	0.136	0.579	0.382	0.125	0.546	0.376	0.122
0.75	0.613	0.454	0.143	0.585	0.407	0.129	0.572	0.402	0.129
(3)									
0.25	0.726	0.563	0.574	0.651	0.522	0.529	0.593	0.491	0.512
0.5	0.742	0.590	0.575	0.667	0.537	0.530	0.671	0.531	0.527
0.75	0.848	0.663	0.308	0.771	0.615	0.253	0.746	0.607	0.267
(4)									
0.25	0.662	0.526	0.177	0.475	0.295	0.220	0.507	0.352	0.155
0.5	0.763	0.551	0.192	0.624	0.434	0.185	0.669	0.483	0.163
0.75	0.941	0.654	0.261	0.816	0.601	0.214	0.775	0.574	0.204

With simulated Poisson data, the Poisson model performs only slightly better than the Discrete Weibull model. Moreover, in almost all the cases the Discrete Weibull model performs better than the Jittering approach.

Negative Binomial data: $Y|X \sim$ Negative Binomial$(\mu(x), \sigma(x))$ with the following link functions:

- Case (1). Linear model for $\mu(x)$, σ constant.

$$\begin{aligned} \log(\mu(x)) &= 1 + 1.5x, \\ \log(\sigma) &= 0.7. \end{aligned}$$

- Case (2A). Linear model for $\mu(x)$, $\sigma(x)$.

$$\begin{aligned} \log(\mu(x)) &= 0.5 + 0.7x \\ \log(\sigma(x)) &= -1 + 0.5x. \end{aligned}$$

- Case (2B). Linear model for $\mu(x)$, $\sigma(x)$, and two covariates affecting different parts of the distribution.

$$\begin{aligned} \log(\mu(x)) &= 0.3 + 0.7x_1 \\ \log(\sigma(x)) &= -2 + 2x_2. \end{aligned}$$

- Case (3). Third degree polynomial model for $\mu(x)$, σ constant.

$$\begin{aligned} \log(\mu(x)) &= 1.5 + x - 0.5x^2 + 0.8x^3, \\ \log(\sigma) &= -2. \end{aligned}$$

- Case (4). Cubic spline model for $\mu(x)$, σ constant.

$$\begin{aligned} \log(\mu(x)) = 1.5 &+ 2x + x^2 + 1.7x^3 + 1.2(x-h_1)^3 I(x > h_1) + \\ &+ 1.4(x-h_2)^3 I(x > h_2) + 2.3(x-h_3)^3 I(x > h_3), \\ \log(\sigma) &= -2. \end{aligned}$$

Table 4.6 reports the square root of the error as in Equation 4.8 for the quantiles $\tau = 0.25, 0.5, 0.75$, averaged over 100 iterations, and for the different sample sizes $n = 50, 100, 1000$.
With simulated Negative Binomial data, the Negative Binomial model performs only slightly better than the Discrete Weibull model, but the Discrete Weibull model always performs better than the Jittering approach. Here Case (2B) is of particular interest: this is the case in which the dispersion depends on a regressor that does not affect the mean. The model estimates are presented in Table 4.7 for the Discrete Weibull, Negative Binomial and Jittering model, respectively. It is interesting to note that the Discrete Weibull model on $q(x)$ and $\beta(x)$ is behaving similarly to the Negative Binomial model, by selecting only x_1 significant in predicting $q(x)$, and only x_2 for $\beta(x)$, while the Jittering approach is able to detect τ-dependent significant variables, which is clearly not possible for a parametric model.

TABLE 4.6: Comparison of different models in terms of root mean squared error on simulated Negative Binomial data under four different model specifications: (1) linear link on q, constant β, (2) linear link on both q and β with (2B) and without (2A) tail behaviour, (3) cubic polynomial link on q, constant β, (4) cubic spline on q, constant β.

	Jittering			DW			NB		
τ\n	50	100	1000	50	100	1000	50	100	1000
(1)									
0.25	0.636	0.493	0.257	0.560	0.428	0.221	0.537	0.401	0.190
0.5	1.564	1.051	0.399	1.193	0.810	0.340	1.200	0.807	0.342
0.75	3.343	2.214	0.656	2.455	1.680	0.580	2.517	1.644	0.577
(2A)									
0.25	0.572	0.480	0.262	0.517	0.419	0.251	0.503	0.416	0.244
0.5	0.651	0.532	0.294	0.593	0.476	0.269	0.580	0.473	0.268
0.75	0.790	0.611	0.318	0.717	0.570	0.306	0.712	0.562	0.302
(2B)									
0.25	0.517	0.505	0.108	0.497	0.477	0.102	0.509	0.480	0.094
0.5	0.538	0.495	0.095	0.514	0.479	0.088	0.555	0.486	0.084
0.75	0.694	0.607	0.143	0.622	0.577	0.122	0.667	0.578	0.119
(3)									
0.25	0.891	0.743	0.571	0.661	0.610	0.537	0.716	0.638	0.536
0.5	0.975	0.720	0.218	0.773	0.613	0.201	0.795	0.633	0.199
0.75	1.224	0.857	0.283	1.054	0.798	0.268	1.031	0.776	0.248
(4)									
0.25	1.173	0.923	0.420	0.847	0.656	0.371	0.927	0.718	0.362
0.5	1.372	0.973	0.502	1.109	0.830	0.452	1.109	0.850	0.443
0.75	1.861	1.192	0.511	1.446	1.056	0.503	1.363	1.007	0.454

TABLE 4.7: Parameter estimates for Discrete Weibull, Negative Binomial and Jittering model from the case (2B) of Negative Binomial simulated data with tail behaviour.

	DW		NB		Jittering		
	$q(x)$	$\beta(x)$	$\mu(x)$	$\sigma(x)$	τ=.25	τ=.5	τ=.75
(Intercept)	0.775***	0.564***	0.354***	-2.219***	-0.136	0.21*	0.636***
	(0.062)	(0.075)	(0.078)	(0.392)	(0.135)	(0.094)	(0.103)
x_1	0.538***	-0.116	0.696***	0.764.	0.434*	0.678***	0.802***
	(0.084)	(0.096)	(0.101)	(0.41)	(0.185)	(0.131)	(0.133)
x_2	-0.094	-0.364***	-0.032	1.466***	-0.503*	-0.219	-0.026
	(0.088)	(0.096)	(0.103)	(0.428)	(0.201)	(0.146)	(0.158)

The coefficients and standard errors (in brackets) are reported.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.6 Real data study

4.6.1 Over-dispersed data

Waiting times before intervention We now continue the evaluation analysis of the healthcare system of Lombardy region in Italy. In chapter 2 we have considered the hospital effectiveness quantifying the hospitals reaction to the newly adopted P4P program. Then, in chapter 3 we investigated the hospital efficiency measuring the patient length of stay on three patient health conditions, i.e. CABG, PTCA, and HIP. Here we measure the hospital efficiency in terms of waiting time before intervention on the three patient health conditions previously considered. Specifically, the waiting times before the intervention is calculated as a difference between the date of the surgery and the booking date. Thus, we consider a total of 16,605 hospitalisations within 109 hospitals, as this is the dimension of the booked surgeries performed for these conditions, and of which 2,487 hospitalisations were for CABG, 6,937 for HIP, and 7,181 for PTCA, respectively. In other words, 70% of the surgeries performed in 2014 for these three health conditions was previously booked. The response variable is measured in weeks, and shows a dispersion of approximately 12, a mean of 11.52, and a range of [0,129]. On average the waiting time for CABG is approximately 3 weeks, while for HIP and PTCA is approximately 13 weeks. Figure 4.4 plots the frequency of the response variable.

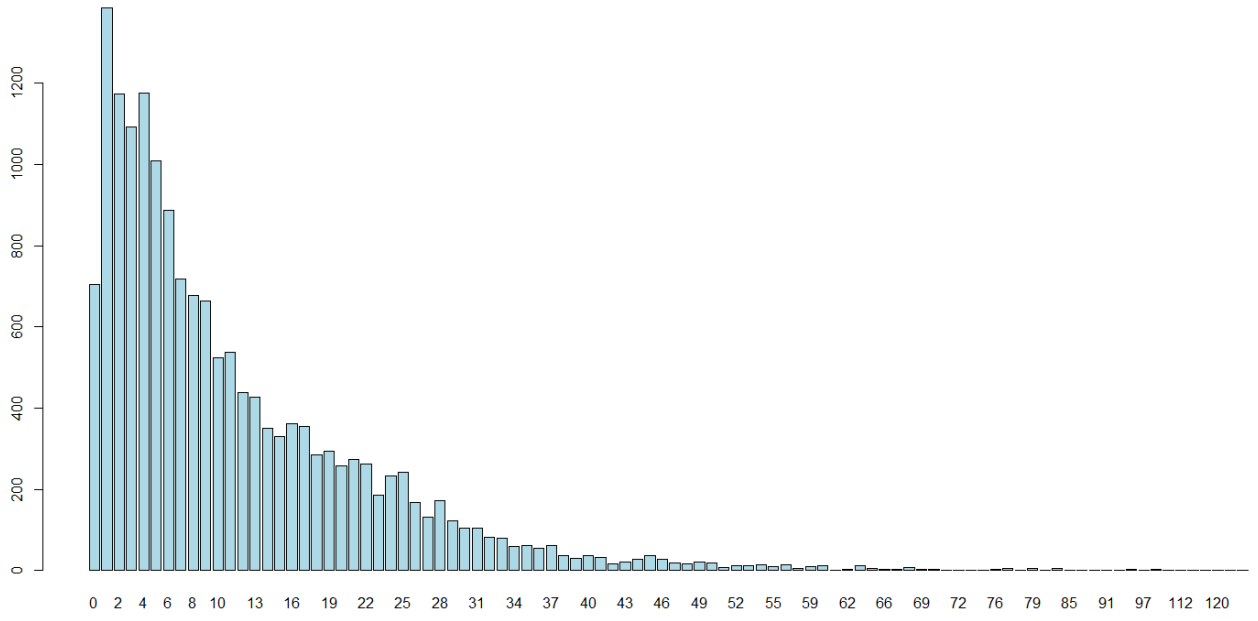


Figure 4.4: Bar plot of the hospital waiting times measured in weeks.

As in section 3.6.1, the structure of the data is calling for a mixed effects model. Thus, we fit a two-level random intercept model considering the patient p (with $p = 1, \dots, P_h$) hospitalised within hospital h (with $h = 1, \dots, H$). Moreover, we include information regarding the patient-reported health procedure, patients' demographic characteristics, i.e. gender and age of the patients, and the comorbidity index. To add flexibility to our model we smooth the data by including a cubic B-splines term with three internal knots set at the quartiles of the variable age of the patient. Thus, recalling the

random intercept model in Equation 3.17, the non-linear mixed effects Discrete Weibull model can be written as

$$\begin{aligned}\log(-\log(q(x,u))) &= \theta_{00} + \theta_{01}\text{female}_{ph} + \theta_{02}\text{age}_{ph} + \theta_{03}\text{age}^2_{ph} + \theta_{04}\text{age}^3_{ph} + \theta_1(\text{age}_{ph}-h_1)^3 I(\text{age}_{ph}>h_1)+\\ &\quad \theta_2(\text{age}_{ph}-h_2)^3 I(\text{age}_{ph}>h_2) + \theta_3(\text{age}_{ph}-h_3)^3 I(\text{age}_{ph}>h_3) + \theta_{05}\text{comorbidity1}_{ph}+\\ &\quad \theta_{06}\text{comorbidity2}_{ph} + \theta_{07}\text{comorbidity3}_{ph} + \theta_{08}\text{procHIP}_{ph} + \theta_{09}\text{procCABG}_{ph} + u_h\\ \log(\beta(x,u)) &= \vartheta_{00} + \vartheta_{01}\text{female}_{ph} + \vartheta_{02}\text{age}_{ph} + \vartheta_{03}\text{age}^2_{ph} + \vartheta_{04}\text{age}^3_{ph} + \vartheta_1(\text{age}_{ph}-h_1)^3 I(\text{age}_{ph}>h_1)+\\ &\quad \vartheta_2(\text{age}_{ph}-h_2)^3 I(\text{age}_{ph}>h_2) + \vartheta_3(\text{age}_{ph}-h_3)^3 I(\text{age}_{ph}>h_3) + \vartheta_{05}\text{comorbidity1}_{ph}+\\ &\quad \vartheta_{06}\text{comorbidity2}_{ph} + \vartheta_{07}\text{comorbidity3}_{ph} + \vartheta_{08}\text{procHIP}_{ph} + \vartheta_{09}\text{procCABG}_{ph} + u_h.\end{aligned} \tag{4.13}$$

As we have previously done in the case of over-dispersed data, we compare our model with the Poisson, Negative Binomial, Poisson-Inverse Gaussian, and COM-Poisson models. The COM-Poisson model for both the parameter $\mu(x)$ and $\sigma(x)$ has some convergence issues, thus we consider the COM-Poisson model with σ constant.

TABLE 4.8: Parameter estimates and AIC values for the non-linear mixed effects models with a cubic B-spline with three internal knots of the variable AGE for the waiting times data.

	PO	CMP	NB		PIG		DW		Jittering		
	$\mu(x)$	$\mu(x)$	$\mu(x)$	$\sigma(x)$	$\mu(x)$	$\sigma(x)$	$q(x)$	$\beta(x)$	τ=.25	τ=.5	τ=.75
Fixed part											
(Intercept)	3.83***	1.176	3.83***	-3.658*	4.18***	-4.47***	3.14***	-0.408	13.62*	19.48**	21.89***
	(0.205)	(0.134)	(0.232)	(1.458)	(0.215)	(1.079)	(0.558)	(0.326)	(5.896)	(5.894)	(6.003)
female	-0.03***	-0.011	-0.031**	-0.021	-0.031**	-0.006	-0.009	0.018	0.017	-0.012	-0.168
	(0.005)	(0.003)	(0.011)	(0.03)	(0.012)	(0.037)	(0.011)	(0.012)	(0.051)	(0.088)	(0.208)
cs.age1	-2.4***	-0.996	-2.39***	3.373	-2.57***	4.98**	-1.356.	0.884.	-8.456	-10.655	-10.39
	(0.302)	(0.195)	(0.385)	(2.097)	(0.377)	(1.832)	(0.756)	(0.503)	(7.795)	(7.824)	(7.669)
cs.age2	-1.01***	-0.426	-1.02***	2.588.	-1.18***	3.44***	-0.339	0.917**	-5.502	-6.63	-3.309
	(0.196)	(0.127)	(0.222)	(1.374)	(0.206)	(0.98)	(0.533)	(0.303)	(5.503)	(5.587)	(5.229)
cs.age3	-1.20***	-0.5	-1.21***	2.829.	-1.36***	3.86***	-0.522	0.797*	-6.038	-7.693	-5.419
	(0.206)	(0.133)	(0.235)	(1.472)	(0.218)	(1.098)	(0.562)	(0.33)	(5.875)	(5.927)	(5.799)
cs.age4	-1.26***	-0.521	-1.27***	2.51.	-1.424***	3.25**	-0.562	0.965**	-5.538	-6.927	-5.621
	(0.203)	(0.131)	(0.232)	(1.439)	(0.216)	(1.057)	(0.553)	(0.321)	(5.754)	(5.834)	(5.601)
cs.age5	-1.52***	-0.594	-1.53***	4.3**	-1.68***	5.99***	-0.809	0.182	-10.97.	-13.303*	-8.98
	(0.223)	(0.143)	(0.293)	(1.547)	(0.288)	(1.244)	(0.598)	(0.382)	(6.033)	(6.36)	(6.632)
cs.age6	-4.11***	-1.617	-4.12***	4.383**	-4.27***	5.32**	-3.11***	-0.137	-14.53*	-12.202	-12.95
	(0.303)	(0.194)	(0.587)	(1.673)	(0.649)	(1.68)	(0.752)	(0.563)	(6.244)	(8.741)	(8.378)
comorbidity1	-0.03**	-0.007	-0.031	-0.054	-0.03	-0.053	-0.019	0.055*	-0.001	0.015	-0.688
	(0.011)	(0.007)	(0.022)	(0.067)	(0.023)	(0.083)	(0.02)	(0.025)	(0.328)	(0.366)	(0.455)
comorbidity2	0.047*	0.021	0.046	0.195.	0.048	0.28.	0.028	-0.096*	0.000	0.902	-0.736
	(0.022)	(0.014)	(0.047)	(0.115)	(0.051)	(0.149)	(0.046)	(0.043)	(0.479)	(0.701)	(1.027)
comorbidity3	-0.31***	-0.125	-0.32***	-1.41**	-0.32***	-1.588**	-0.28***	0.43***	0.01	0.071	-3.42**
	(0.064)	(0.042)	(0.079)	(0.49)	(0.08)	(0.507)	(0.065)	(0.09)	(0.476)	(0.604)	(1.2)
procHIP	0.02***	0.007	0.02.	-0.11***	0.02.	-0.13***	0.022*	0.04**	0.034	0.049	0.127*
	(0.005)	(0.003)	(0.011)	(0.03)	(0.012)	(0.037)	(0.011)	(0.013)	(0.099)	(0.065)	(0.047)
procCABG	-1.78***	-0.693	-1.78***	0.24***	-1.79***	0.293***	-1.62***	-0.22***	-6.68***	-10.9***	-14.91***
	(0.014)	(0.01)	(0.024)	(0.059)	(0.025)	(0.075)	(0.022)	(0.021)	(1.472)	(1.803)	(2.569)
Random part											
var(u_{0j})	0.293	0.041	0.293		0.311		0.260		5.122	14.59	24.12
AIC	155565.6	111852.5	104407.3		104417.3		104362.4		-	-	-
logLik	-77665.48	-56154.53	-51978.65		-51977.91		-51960.44		-	-	-

The coefficients and standard errors (in brackets) are reported.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.8 reports the parameter estimates and the AIC of these models, which point to the choice of the Discrete Weibull as the best fitting model. We note that the AIC value for the linear mixed effects Discrete Weibull model, i.e. fitted without the cubic B-spline for age, is higher than the AIC value of the non-linear mixed effects model, i.e. AIC linear: 104,556. Figure 4.5 shows the diagnostic plot of the theoretical versus the sample quantiles.

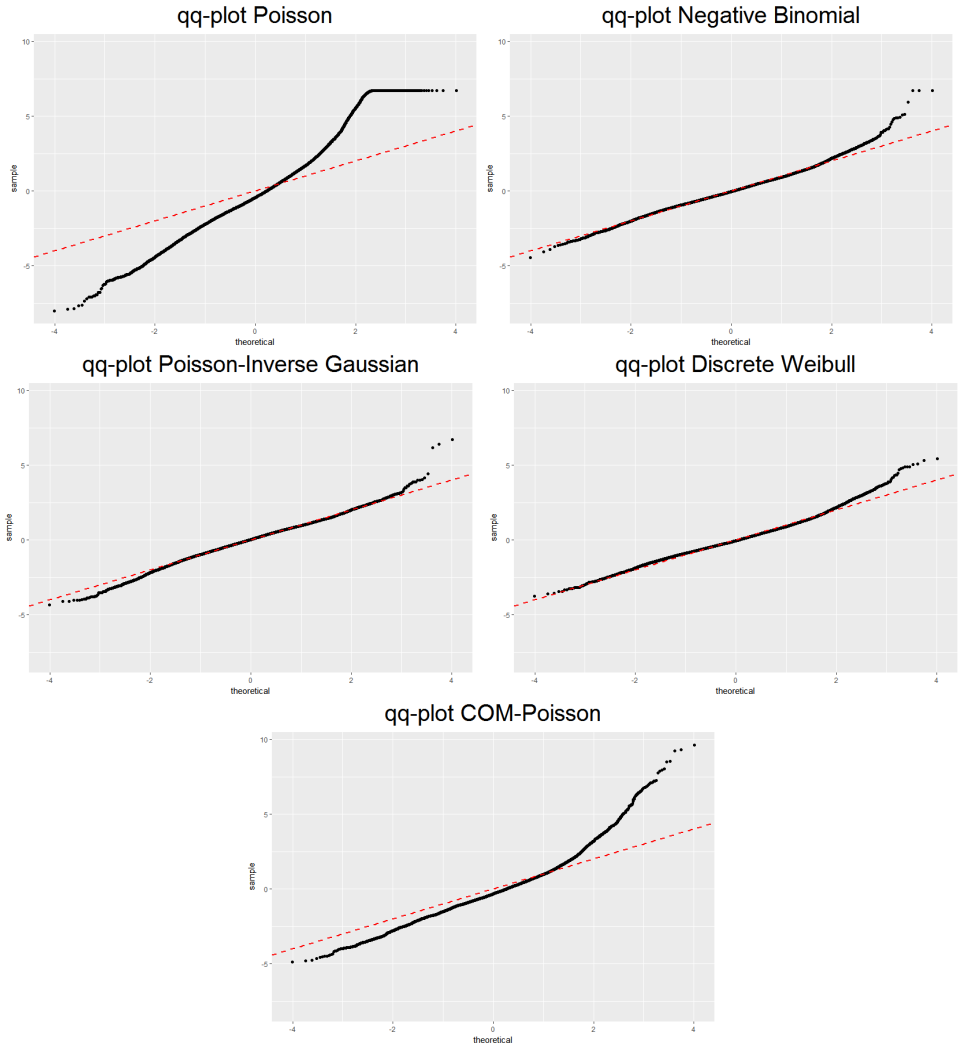


FIGURE 4.5: Diagnostic plots for the analyses of the waiting times data using various non-linear mixed effects regression models.

We note that being admitted for CABG has a strong negative effect on the waiting times, particularly on the upper tail. This coefficient is found significant both from the Discrete Weibull model and the Jittering approach at the 25%, 50% and 75%. The sex of the patient does not impact the waiting times. Patients with many comorbidities and older patients present decreasing waiting times. Overall this analysis shows how the Discrete Weibull distribution represents a competitive alternatives to available parametric regression models for over-dispersed data. Moreover, the non-linear Discrete Weibull model has a comparable performance to the more complex Jittering approach and allows to detect similar dependencies. In addition to this, the main theoretical advantage of our parametric approach via a Discrete Weibull distribution on modelling this data is the non-crossing quantiles as showed in section 1.4 for these data.

Unnecessary hospital bed occupancy We model data from [41], which are available in the `R` package `gamlss.data` under the name aep. This study was carried out at the Hospital del Mar in Barcelona (Spain) during the years 1988 and 1990. The aim of the study is to model the number of inappropriate days out of the total number of days spent in hospital.

In particular, a reduction of the inappropriate stays in hospital could increase the hospital productivity and reduce the waiting lists. The response is regressed against the type of ward in the hospital as a factor with three categories, i.e. medical, surgical, and others, the specific year as a factor with two categories, i.e. year 1988 and 1990, and the gender of the patient. Thus, we consider n=620 observations, where the response variable has a mean of 7.23, a variance of 72.52, and a range of [1,72]. The dispersion for the response variable is close to 10, thus we are modelling over-dispersed data. Figure 4.6 plots the frequencies of the response variable.

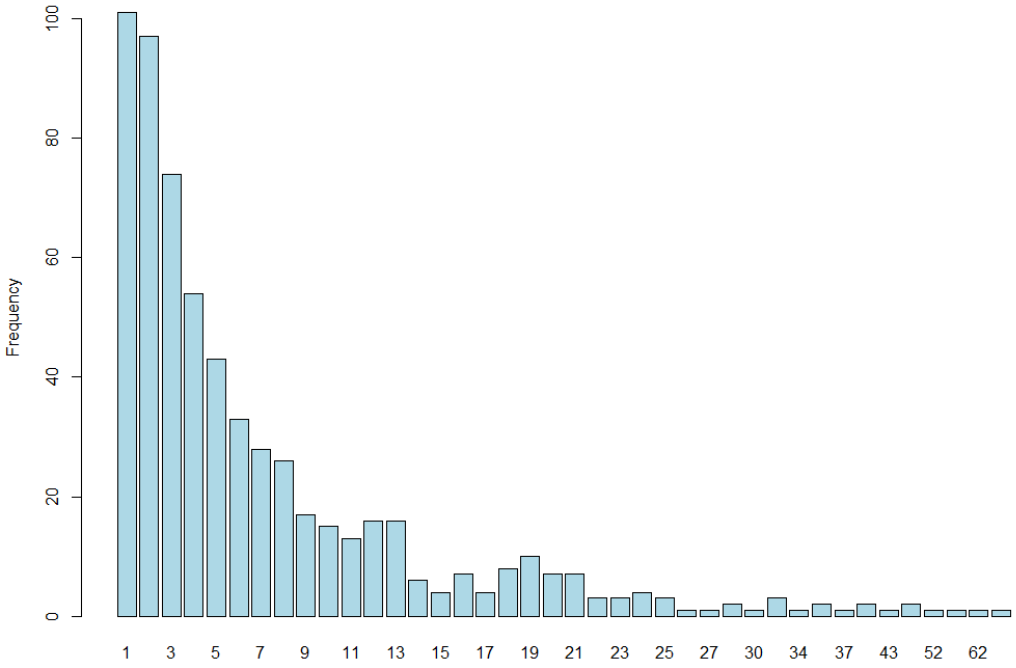


FIGURE 4.6: Bar plot of the unnecessary hospital bed occupancy measured in days.

We perform a Discrete Weibull linear regression model for both $q(x)$ and $\beta(x)$, i.e.

$$\begin{aligned} \log(-\log(q(x))) &= \theta_{00} + \theta_{01}\text{age} + \theta_{02}\text{los} + \theta_{03}\text{wardS} + \\ &\quad \theta_{04}\text{wardO} + \theta_{05}\text{year90} + \theta_{06}\text{female} \\ \log(\beta(x)) &= \vartheta_{00} + \vartheta_{01}\text{age} + \vartheta_{02}\text{los} + \vartheta_{03}\text{wardS} + \\ &\quad \vartheta_{04}\text{wardO} + \vartheta_{05}\text{year90} + \vartheta_{06}\text{female}. \end{aligned} \tag{4.14}$$

Table 4.9 shows the parameter estimates via different parametric regression model specifications and for the Jittering approach for the 25%, 50% and 75% quantiles. The COM-Poisson model for both the parameter $\mu(x)$ and $\sigma(x)$ has some convergence issues, thus in Table 4.9 is presented the COM-Poisson model with σ constant.

TABLE 4.9: Parameter estimates for different parametric regression model specifications and the Jittering approach for the unnecessary hospital bed occupancy data.

	PO	CMP	NB		PIG		DW		Jittering		
	$\mu(x)$	$\mu(x)$	$\mu(x)$	$\sigma(x)$	$\mu(x)$	$\sigma(x)$	$q(x)$	$\beta(x)$	τ=.25	τ=.5	τ=.75
(Intercept)	1.478***	0.27**	1.089***	-2.791***	0.971***	-3.312***	1.322***	1.052***	0.796***	1.173***	1.486***
	(0.033)	(0.094)	(0.068)	(0.284)	(0.074)	(0.338)	(0.057)	(0.079)	(0.075)	(0.101)	(0.104)
age	0.007***	0.003	0.003*	-0.005	0.002	-0.009	0.003*	0.001	0.004	0.004**	0.004
	(0.001)	(0.044)	(0.001)	(0.006)	(0.002)	(0.007)	(0.001)	(0.002)	(0.003)	(0.002)	(0.003)
los	0.033***	0.011***	0.062***	0.044***	0.072***	0.081***	0.058***	-0.015***	0.046***	0.053***	0.055***
	(0.001)	(0.001)	(0.004)	(0.007)	(0.005)	(0.01)	(0.003)	(0.003)	(0.002)	(0.004)	(0.006)
wardS	-0.328***	-0.132***	-0.38***	0.291	-0.378***	0.293	-0.357***	-0.143*	-0.419***	-0.571***	-0.48***
	(0.031)	(0.001)	(0.052)	(0.218)	(0.053)	(0.26)	(0.045)	(0.064)	(0.072)	(0.072)	(0.109)
wardO	-0.412**	-0.149***	-0.423*	0.132	-0.454*	-0.123	-0.376*	-0.083	-0.645*	-0.508	-0.257
	(0.142)	(0.02)	(0.2)	(0.919)	(0.197)	(1.265)	(0.165)	(0.236)	(0.286)	(0.428)	(0.252)
year90	-0.239***	-0.085	-0.224***	0.676**	-0.202***	0.796**	-0.206***	-0.258***	-0.402***	-0.386***	-0.261**
	(0.032)	(0.09)	(0.052)	(0.225)	(0.054)	(0.265)	(0.045)	(0.065)	(0.073)	(0.072)	(0.084)
female	0.046	0.005	-0.025	-0.141	-0.034	-0.244	-0.023	0.022	0.007	-0.023	-0.031
	(0.031)	(0.019)	(0.05)	(0.225)	(0.052)	(0.265)	(0.043)	(0.065)	(0.073)	(0.075)	(0.092)
AIC	3903.841	3354.22	3184.352		3164.867		3156.272		-	-	-

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We note that both the parametric models via different distributions and the Jittering, detect the same variables as significant predictors. In terms of AIC, the Discrete Weibull and the Poisson-Inverse Gaussian models lead to the best fit. The variance ratio plot for these models in Figure 4.7 confirms these results, as well as the diagnostic plot in Figure 4.8.

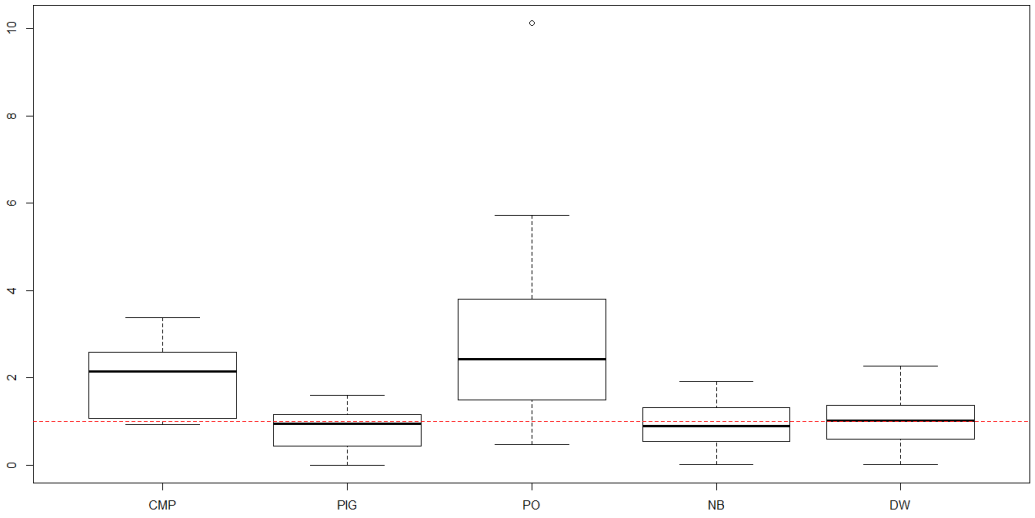


FIGURE 4.7: Variance ratio plots of five different models for the unnecessary hospital bed occupancy data.

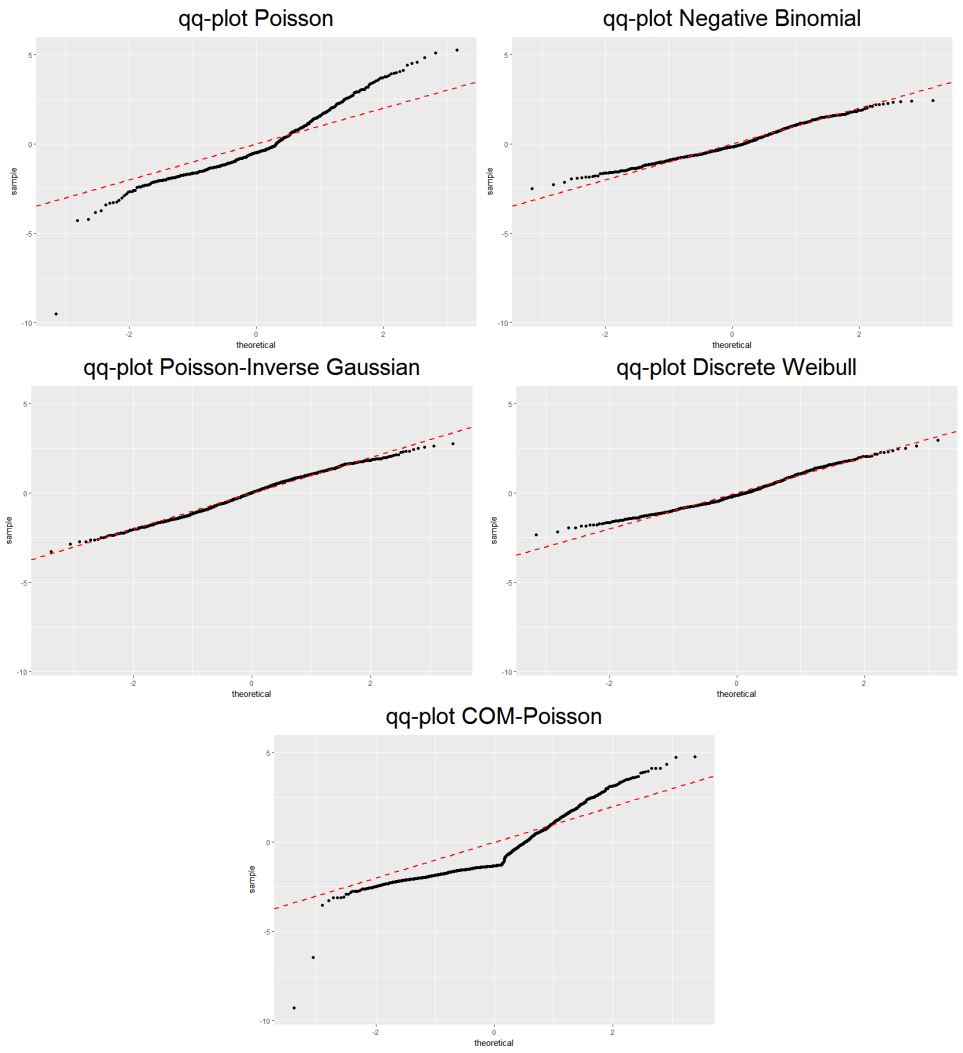


FIGURE 4.8: Diagnostic plots for the linear analyses over both the distributional parameters of the unnecessary hospital bed occupancy data using various regression models.

We now assess the goodness-of-fit of the Discrete Weibull model by comparing the observed and expected number of data points in each of the ten regions defined by the 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% quantiles.

TABLE 4.10: Observed and expected number of data points by region for the Jittering and Discrete Weibull model on the unnecessary hospital bed occupancy data.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$n^{(\tau)}$	62	62	62	62	62	62	62	62	62	62
$n^{(\tau)}$ DW	93	107	69	65	50	42	39	47	45	63
$n^{(\tau)}$ Jittering	122	71	46	62	61	50	48	55	60	45
$n^{(\tau)}$ (%)	10	10	10	10	10	10	10	10	10	10
$n^{(\tau)}$ DW (%)	15	17.26	11.13	10.48	8.06	6.77	6.29	7.58	7.26	10.16
$n^{(\tau)}$ Jittering (%)	19.68	11.45	7.42	10	9.84	8.06	7.74	8.87	9.68	7.26

Table 4.10 reports the number of observations in each region. We would expect 62 observations (10%) in each of the regions. The numbers of observation obtained with the two approaches are very similar. This is shown visually in Figure 4.9, which plots the expected percentage of data points in each regions for both the Discrete Weibull (red) and the Jittering (blue) model.

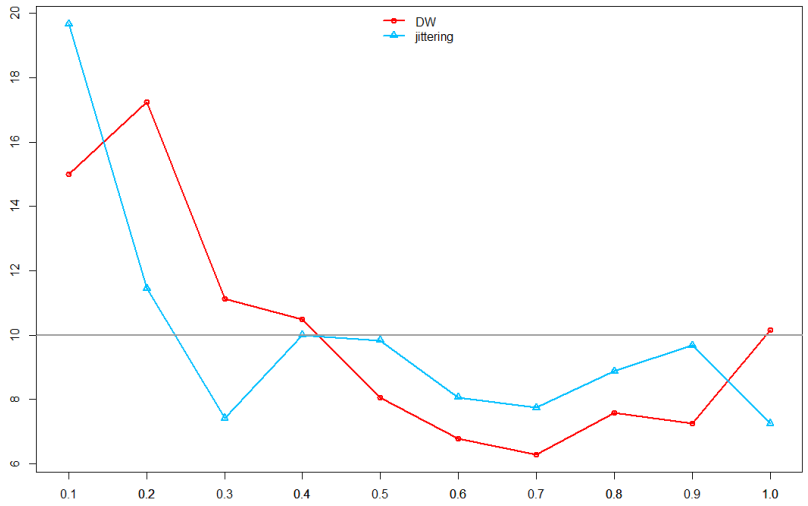


FIGURE 4.9: Expected percentage of data points (y-axis) for each τ-quantiles and by region for the Jittering and Discrete Weibull model on the unnecessary hospital bed occupancy data.

We now focus on the estimation of the partial effects in order to quantify the change in the quantiles of the dependent variable in response to a change in each explanatory variable, while keeping all the other covariates constant as described in section 4.4. Table 4.11 reports these partial effects. We can conclude that there are similarities regarding the intensity and the signs of the effects between the two models.

TABLE 4.11: Partial effects for the Discrete Weibull and Jittering models on the unnecessary hospital bed occupancy data.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Jittering								
age	0.001	0.002	0.003	0.005	0.013	0.032	0.058	0.061	0.026
los	0.004	0.011	0.020	0.037	0.079	0.194	0.341	0.590	0.809
wardS	-0.022	-0.040	-0.045	-0.099	-0.351	-1.055	-1.551	-2.712	-3.072
wardO	-0.071	-0.134	-0.167	-0.264	-0.576	-1.409	-2.330	-4.301	-3.978
year90	0.060	0.150	0.227	0.325	0.481	0.461	0.107	-0.734	-0.211
female	-0.018	-0.027	-0.039	-0.051	-0.042	-0.250	-0.638	-0.840	-0.938
	Discrete Weibull								
age	0.001	0.003	0.005	0.008	0.011	0.016	0.021	0.027	0.036
los	0.004	0.015	0.036	0.069	0.120	0.198	0.321	0.536	1.004
wardS	-0.005	-0.035	-0.095	-0.194	-0.348	-0.587	-0.969	-1.638	-3.098
wardO	-0.024	-0.088	-0.195	-0.357	-0.596	-0.950	-1.496	-2.421	-4.373
year90	0.142	0.289	0.429	0.555	0.657	0.716	0.694	0.489	-0.292
female	-0.008	-0.022	-0.044	-0.075	-0.116	-0.174	-0.257	-0.389	-0.649

This example shows a very good fit of the linear Discrete Weibull model where both the parameters q and β are linked to x. Nevertheless, the performance of this model can improve by considering a more local approach. Specifically, we now

fit a linear regression Discrete Weibull model with Gaussian kernel weights for $q(x)$ and $\beta(x)$, which can be written as

$$\begin{aligned}\log\left(-\log\left((q(x))^{(b)}\right)\right) &= \theta_{00}^{(b)} + \theta_{01}^{(b)}\text{age} + \theta_{02}^{(b)}\text{los} + \theta_{03}^{(b)}\text{wardS} + \\ &\quad \theta_{04}^{(b)}\text{wardO} + \theta_{05}^{(b)}\text{year90} + \theta_{06}^{(b)}\text{female} \\ \log\left((\beta(x))^{(b)}\right) &= \vartheta_{00}^{(b)} + \vartheta_{01}^{(b)}\text{age} + \vartheta_{02}^{(b)}\text{los} + \vartheta_{03}^{(b)}\text{wardS} + \\ &\quad \vartheta_{04}^{(b)}\text{wardO} + \vartheta_{05}^{(b)}\text{year90} + \vartheta_{06}^{(b)}\text{female},\end{aligned} \tag{4.15}$$

where estimates of the parameters are obtained for each bandwidth b considered.

TABLE 4.12: Parameter estimates and AIC values for the linear Discrete Weibull($q(x)$,$\beta(x)$) model with Gaussian kernel weights set at the bandwidth b=(0.2,0.1,0.001) for the unnecessary hospital bed occupancy data.

	KLR b=0.2 DW		KLR b=0.1 DW		KLR b=0.001 DW	
	$q(x)$	$\beta(x)$	$q(x)$	$\beta(x)$	$q(x)$	$\beta(x)$
(Intercept)	1.308***	1.058***	1.319***	1.054***	1.322***	1.052***
	(0.061)	(0.085)	(0.058)	(0.081)	(0.057)	(0.079)
age	0.002.	0.002	0.003*	0.002	0.003*	0.001
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
los	0.06***	-0.015***	0.059***	-0.015***	0.058***	-0.015***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
wardS	-0.355***	-0.143*	-0.357***	-0.143*	-0.357***	-0.143*
	(0.048)	(0.07)	(0.045)	(0.066)	(0.045)	(0.064)
wardO	-0.37	-0.081	-0.374*	-0.083	-0.376*	-0.083
	(0.276)	(0.397)	(0.188)	(0.268)	(0.165)	(0.236)
year90	-0.205***	-0.266***	-0.206***	-0.26***	-0.206***	-0.258***
	(0.048)	(0.07)	(0.046)	(0.066)	(0.045)	(0.065)
female	-0.023	0.02	-0.023	0.021	-0.023	0.022
	(0.046)	(0.07)	(0.044)	(0.066)	(0.043)	(0.065)
AIC	2674.269		3021.552		3156.258	

The coefficients and standard errors (in brackets) are reported.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.12 show how the parameters estimates for the cases when the bandwidth is set to 0.2, 0.1 and 0.001 respectively. We note that for large values of the bandwidth the estimation becomes more local, while for small values of the bandwidth, i.e. b=0.001, the estimates reduce to the maximum likelihood estimates presented in Table 4.9. We now check whether the partial effects of this model could get closer to the ones of the Jittering approach. Table 4.13 reports the partial effects for the three chosen bandwidth. To deeper investigate the fit of the the significant covariates AGE, LOS, and YEAR90, in Figure 4.10, Figure 4.11, and Figure 4.12 we propose a graphical approach to visualise the partial effects for each τ-quantiles and by bandwidth. We note that, depending on the quantiles, for some covariates the local estimator leads to predictions closer to the ones obtained via the Jittering approach.

TABLE 4.13: Partial effects of the regressors on the dependent variable for the Discrete Weibull model with Gaussian kernel weights and where both parameters are linked to the covariates. We report the effects corresponding to the bandwidth b=(0.2,0.1,0.001) for the unnecessary hospital bed occupancy.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Kernel linear regression with b=0.2 Discrete Weibull								
age	0.004	0.018	0.032	0.046	0.061	0.078	0.097	0.122	0.159
los	0.16	0.26	0.349	0.435	0.523	0.617	0.725	0.857	1.053
wardS	-1.421	-1.794	-2.061	-2.282	-2.48	-2.669	-2.858	-3.061	-3.312
wardO	-1.331	-1.738	-2.047	-2.318	-2.573	-2.829	-3.1	-3.414	-3.84
year90	-1.349	-1.561	-1.661	-1.705	-1.709	-1.676	-1.6	-1.458	-1.17
female	-0.031	-0.067	-0.099	-0.131	-0.164	-0.201	-0.243	-0.295	-0.374
	Kernel linear regression with b=0.1 Discrete Weibull								
age	0.012	0.017	0.021	0.026	0.03	0.034	0.039	0.045	0.054
los	0.159	0.247	0.323	0.396	0.469	0.548	0.637	0.747	0.908
wardS	-1.4	-1.78	-2.055	-2.286	-2.495	-2.697	-2.902	-3.127	-3.412
wardO	-1.301	-1.707	-2.018	-2.291	-2.549	-2.809	-3.087	-3.409	-3.849
year90	-1.358	-1.578	-1.686	-1.737	-1.747	-1.722	-1.653	-1.519	-1.243
female	-0.012	-0.049	-0.083	-0.118	-0.155	-0.196	-0.244	-0.304	-0.394
	Kernel linear regression with b=0.001 Discrete Weibull								
age	0.014	0.017	0.019	0.021	0.022	0.023	0.025	0.026	0.027
los	0.158	0.243	0.315	0.384	0.454	0.529	0.613	0.717	0.868
wardS	-1.392	-1.773	-2.051	-2.285	-2.498	-2.703	-2.913	-3.144	-3.439
wardO	-1.292	-1.697	-2.008	-2.282	-2.541	-2.802	-3.082	-3.406	-3.849
year90	-1.36	-1.582	-1.692	-1.745	-1.758	-1.735	-1.669	-1.539	-1.267
female	-0.008	-0.044	-0.08	-0.116	-0.154	-0.196	-0.245	-0.307	-0.4

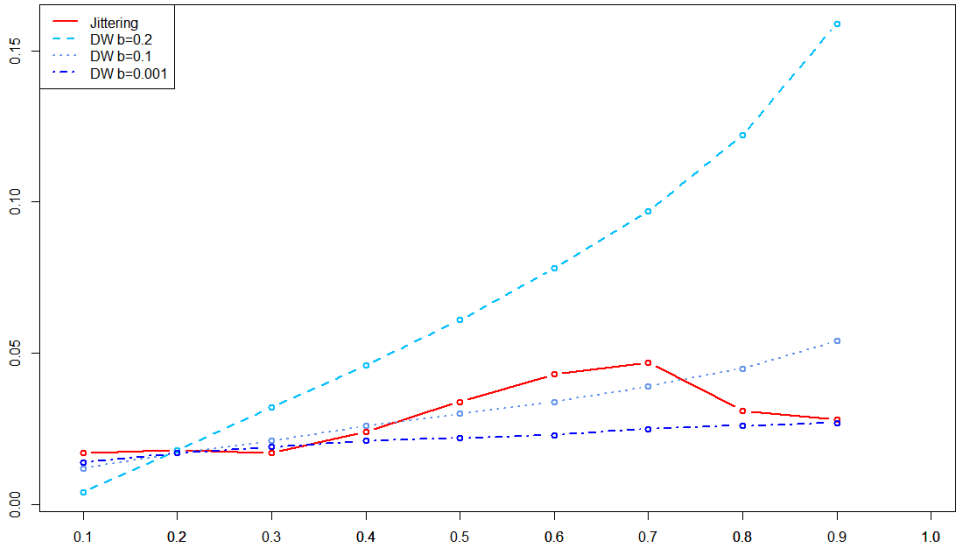


FIGURE 4.10: Partial effects by bandwidth (y-axis) and τ (x-axis) for the variable AGE in the unnecessary hospital bed occupancy data.

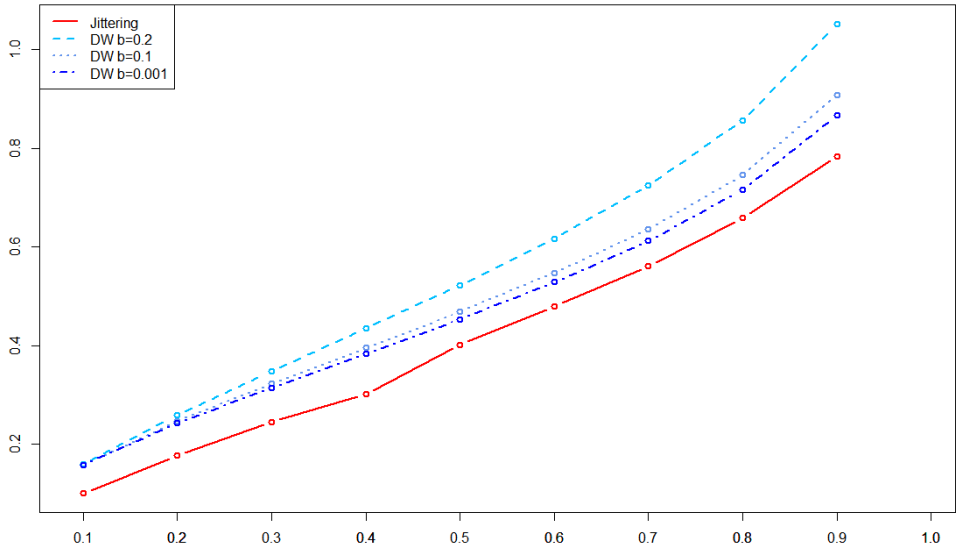


FIGURE 4.11: Partial effects by bandwidth (y-axis) and τ (x-axis) for the variable LOS in the unnecessary hospital bed occupancy data.

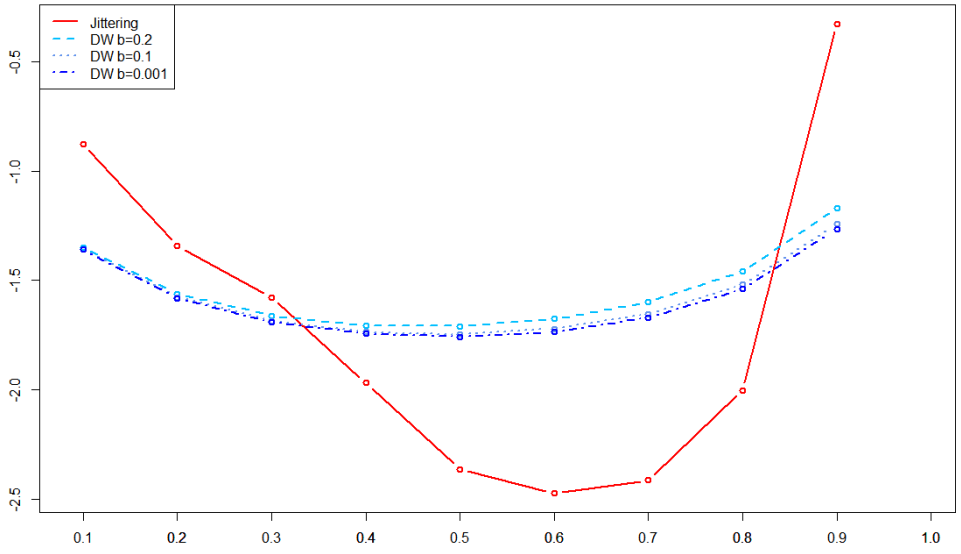


FIGURE 4.12: Partial effects by bandwidth (y-axis) and τ (x-axis) for the variable YEAR90 in the unnecessary hospital bed occupancy data.

4.6.2 Under-dispersed data

Ideal fertility We model data from [69]. The data contain 5,628 observations, from the National Survey of Demographic Dynamics 1997, i.e. ENADID from its acronym in Spanish. In particular, data are collected for women aged between 15 and 17 who at the time of the ENADID interview were living with at least one biological parent and had neither started independent economic life nor entered motherhood. The aim of the study is to examine how education and family background affect the planned fertility of young individuals in Mexico, hence we model the desired number of children. By construction, the sample is composed of women aged between 15 and 17 years old. For this reason, age has not enough variation in the data and will not be considered as an explanatory variable. Thus, as explanatory variable we consider the teenager's number of siblings, whether the she can speak an indigenous/native language, whether she is of catholic religion, and a set of dummies control for the teenager highest education attainment, i.e. incomplete primary, complete primary, incomplete secondary, complete secondary and over secondary. The study also controls for the location of the parental household. Three categories are considered: rural, urban, and suburban. Family background is controlled by a set of variables describing the socio-economic characteristics of the head of the family as his/her age and income, and her/his higher education attainment composed of five categories as for the education of the teenager considered. The family type which reflects the presence of both parents, an absent mother or an absent father, and a series of dummies indicating the birthplace of the teenager are also used as explanatory variables. The mean of the response is 2.5, the variance is 1.37, and the range is [0,12]. The dispersion for the response variable is close to 0.55. Thus, we employ a Poisson, a Generalised Poisson and a COM-Poisson as a comparison. The frequency of the planned fertility are showed in Figure 4.13.

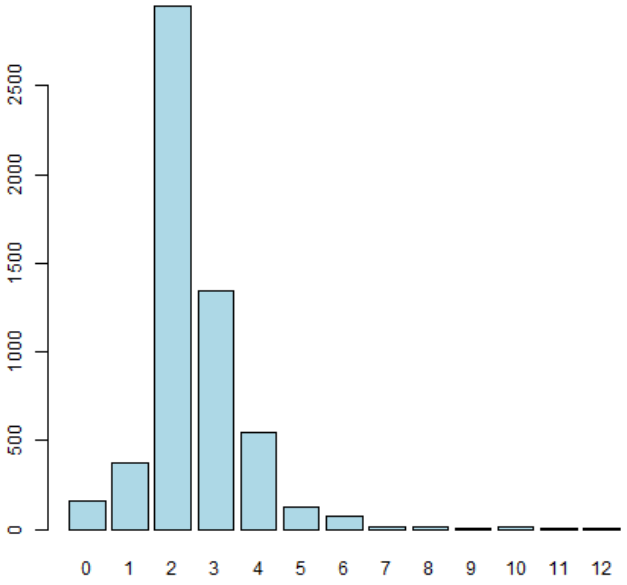


FIGURE 4.13: Bar plot of the planned fertility measured as ideal number of children.

To model these data, [69] employed a non-parametric linear regression model via the Jittering approach. For a fair comparison, we employ the same specification model, i.e. a linear regression Discrete Weibull model for $q(x)$ and $\beta(x)$, which can be written as

$$
\begin{aligned}
\log(-\log(q(x))) &= \theta_0 + \theta_1 \text{siblings} + \theta_2 \text{HFage} + \theta_3 \text{cprimary} + \theta_4 \text{isecondary} + \\
&\quad \theta_5 \text{csecondary} + \theta_6 \text{osecondary} + \theta_7 \text{HFcprimary} + \theta_8 \text{HFisecondary} + \\
&\quad \theta_9 \text{HFcsecondary} + \theta_{10} \text{HFosecondary} + \theta_{11} \text{HFincome} + \theta_{12} \text{catholic} + \\
&\quad \theta_{13} \text{indspker} + \theta_{14} \text{urban} + \theta_{15} \text{surban} + \theta_{16} \text{AbsentFather} + \theta_{17} \text{AbsentMother} + \\
&\quad 32 \text{ birth place dummies} \\
\log(\beta(x)) &= \vartheta_0 + \vartheta_1 \text{siblings} + \vartheta_2 \text{HFage} + \vartheta_3 \text{cprimary} + \vartheta_4 \text{isecondary} + \\
&\quad \vartheta_5 \text{csecondary} + \vartheta_6 \text{osecondary} + \vartheta_7 \text{HFcprimary} + \vartheta_8 \text{HFisecondary} + \\
&\quad \vartheta_9 \text{HFcsecondary} + \vartheta_{10} \text{HFosecondary} + \vartheta_{11} \text{HFincome} + \vartheta_{12} \text{catholic} + \\
&\quad \vartheta_{13} \text{indspker} + \vartheta_{14} \text{urban} + \vartheta_{15} \text{surban} + \vartheta_{16} \text{AbsentFather} + \vartheta_{17} \text{AbsentMother} + \\
&\quad 32 \text{ birth place dummies}
\end{aligned}
\tag{4.16}
$$

By using our parametric approach via the Discrete Weibull model we note the computational gain in terms of CPU time in seconds in a comparison with the COM-Poisson model and the Jittering method averaged over 50 dithered samples and for 9 quantiles, as reported in Table 4.14.

TABLE 4.14: System time (in seconds) performance comparison between the same specification model via the Poisson, the generalised Poisson, the Discrete Weibull, the COM-Poisson distributions, and the Jittering model averaged over 50 dithered samples and for 9 quantiles.

Model	PO	GPO	DW	Jittering	CMP
CPU time	0.14	1.13	3.70	53.94	6,212.37

Table 4.15 shows the parameter estimates via different parametric regression model specifications, and for the Jittering approach, for the 25%, 50% and 75% quantiles. The Generalised-Poisson model for both parameters i.e. $\mu(x)$ and $\sigma(x)$, has some convergence issues, thus in Table 4.15 we present the Generalised-Poisson model with σ constant. The table shows various significant variables picked both by the Jittering and the Discrete Weibull model both in the $q(x)$ and $\beta(x)$ regression part. For this data, in terms of AIC, the COM-Poisson represents the best parametric alternative to the Discrete Weibull model.

Table 4.16 reports the partial effects of the Jittering and the Discrete Weibull model. Considering both the intensity and the sign of the coefficients, we conclude that the effects estimates obtained via the Discrete Weibull model are very similar to the ones obtained via the Jittering approach and showed in [69]. In fact, both approaches revealed effects mostly at the tails of the conditional distribution, e.g. for the education factors.

TABLE 4.15: Parameter estimates and AIC values for the Discrete Weibull model of Equation 4.16 and Jittering on the planned fertility data.

	PO	GPO	CMP		DW		JITTERING		
	$\mu(x)$	$\mu(x)$	$\mu(x)$	$\sigma(x)$	$q(x)$	$\beta(x)$	τ=.25	τ=.5	τ=.75
(Intercept)	0.735***	0.94***	0.687	0.263	1.04***	1.242***	0.608***	0.735***	0.905***
	(0.087)	(0.067)	(0.435)	(0.157)	(0.045)	(0.101)	(0.045)	(0.054)	(0.064)
siblings	0.025***	0.032***	-0.145	-0.073	0.021***	-0.039***	0.01***	0.023***	0.031***
	(0.005)	(0.004)	(0.022)	(0.009)	(0.003)	(0.006)	(0.003)	(0.004)	(0.004)
HFage	0.002	0.003***	0.043	0.014	0.002**	-0.006***	0.001	0.001	0.002*
	(0.001)	(0.001)	(0.007)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
cprimary	0.002	-0.028	-0.078	0.004	0.024	0.114*	0.047.	0.001	-0.026
	(0.037)	(0.028)	(0.172)	(0.067)	(0.026)	(0.046)	(0.028)	(0.036)	(0.036)
isecondary	-0.051	-0.08**	-0.07	0.064	-0.043.	0.199***	0.021	-0.052	-0.087*
	(0.039)	(0.029)	(0.182)	(0.07)	(0.026)	(0.047)	(0.027)	(0.035)	(0.037)
csecondary	-0.061.	-0.098***	-0.004	0.107	-0.05*	0.266***	0.035	-0.061.	-0.128***
	(0.036)	(0.027)	(0.173)	(0.065)	(0.026)	(0.044)	(0.026)	(0.033)	(0.034)
osecondary	-0.073.	-0.125***	-0.006	0.121	-0.066**	0.352***	0.028	-0.072*	-0.131***
	(0.038)	(0.029)	(0.189)	(0.07)	(0.026)	(0.047)	(0.026)	(0.033)	(0.035)
HFcprimary	-0.017	-0.017	0.014	0.038	-0.006	-0.021	-0.007	-0.018	-0.027
	(0.023)	(0.018)	(0.117)	(0.043)	(0.026)	(0.027)	(0.011)	(0.014)	(0.018)
HFisecondary	-0.013	-0.03	-0.013	0.036	-0.012	0.095.	-0.006	-0.013	-0.001
	(0.049)	(0.038)	(0.262)	(0.096)	(0.026)	(0.057)	(0.02)	(0.025)	(0.04)
HFcsecondary	-0.056.	-0.072**	-0.02	0.071	-0.042**	0.109**	-0.016	-0.056***	-0.074***
	(0.033)	(0.025)	(0.171)	(0.062)	(0.026)	(0.037)	(0.013)	(0.015)	(0.022)
HFosecondary	-0.059.	-0.06*	-0.031	0.073	-0.043**	-0.021	-0.03*	-0.058***	-0.059**
	(0.031)	(0.024)	(0.169)	(0.058)	(0.026)	(0.036)	(0.013)	(0.015)	(0.021)
HFincome	0.005.	0.006**	0.063	0.015	0.004**	-0.005	0.003.	0.006***	0.004*
	(0.003)	(0.002)	(0.014)	(0.003)	(0.026)	(0.003)	(0.002)	(0.001)	(0.002)
catholic	0.024	0.011	0.074	0.018	0.016	0.009	0.013	0.024	0.01
	(0.03)	(0.022)	(0.14)	(0.052)	(0.026)	(0.034)	(0.014)	(0.017)	(0.02)
indspker	-0.001	0.059.	-0.05	-0.057	0.023	-0.303***	-0.063.	-0.001	0.063
	(0.044)	(0.033)	(0.204)	(0.081)	(0.026)	(0.053)	(0.034)	(0.037)	(0.043)
urban	-0.11***	-0.132***	0.027	0.109	-0.086***	0.049.	-0.057***	-0.109***	-0.14***
	(0.023)	(0.017)	(0.12)	(0.044)	(0.026)	(0.027)	(0.011)	(0.014)	(0.018)
surban	-0.06*	-0.062**	-0.084	-0.001	-0.05***	-0.046	-0.047***	-0.061***	-0.077***
	(0.026)	(0.02)	(0.127)	(0.048)	(0.026)	(0.031)	(0.014)	(0.018)	(0.021)
AbsentFather	-0.028	-0.018	-0.025	0.011	-0.007	-0.046	-0.018	-0.028*	-0.026
	(0.026)	(0.02)	(0.136)	(0.049)	(0.026)	(0.029)	(0.011)	(0.013)	(0.018)
AbsentMother	-0.04	-0.07	-0.044	0.025	-0.031	0.117.	0.005	-0.04	-0.045
	(0.056)	(0.043)	(0.288)	(0.099)	(0.026)	(0.064)	(0.025)	(0.03)	(0.035)
AIC	18047.1	17118.9	16468.79		16415.76		-	-	-

Birthplace dummies are included in the regression.

The coefficients and standard errors (in brackets) are reported.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 4.16: Partial effects of the regressors on the dependent variable for the linear Discrete Weibull model where both the distributional parameters are linked to the covariates and Jittering models for the planned fertility data.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
					Jittering				
catholic	0.063	0.030	0.021	0.015	0.014	0.016	0.015	0.056	0.058
indspker	-0.657	-0.145	-0.085	-0.041	0.002	0.066	0.131	0.203	0.540
cprimary	0.301	0.107	0.074	0.049	0.021	-0.027	-0.061	-0.089	-0.194
isecondary	0.295	0.071	0.002	-0.049	-0.100	-0.178	-0.213	-0.230	-0.388
csecondary	0.340	0.107	0.030	-0.024	-0.091	-0.184	-0.273	-0.365	-0.594
osecondary	0.328	0.094	0.021	-0.036	-0.103	-0.196	-0.282	-0.368	-0.602
siblings	0.004	0.013	0.025	0.037	0.051	0.071	0.082	0.081	0.087
urban	-0.043	-0.089	-0.125	-0.177	-0.223	-0.280	-0.328	-0.353	-0.398
surban	-0.053	-0.078	-0.093	-0.123	-0.142	-0.178	-0.203	-0.199	-0.199
HFage	0.001	0.001	0.001	0.001	0.001	0.002	0.004	0.006	0.009
HFcprimary	-0.012	-0.013	-0.015	-0.022	-0.036	-0.055	-0.064	-0.077	-0.075
HFisecondary	-0.010	-0.016	-0.010	-0.006	-0.020	-0.035	-0.024	0.010	0.047
HFcsecondary	-0.022	-0.023	-0.035	-0.052	-0.079	-0.121	-0.159	-0.210	-0.272
HFosecondary	-0.063	-0.053	-0.059	-0.067	-0.083	-0.105	-0.123	-0.175	-0.226
HFincome	0.004	0.004	0.006	0.008	0.009	0.010	0.011	0.011	0.011
AbsentFather	-0.042	-0.025	-0.034	-0.038	-0.046	-0.054	-0.061	-0.067	-0.068
AbsentMother	0.012	0.021	0.012	0.003	-0.018	-0.038	-0.088	-0.133	-0.196
					Discrete Weibull				
catholic	0.030	0.036	0.041	0.044	0.047	0.050	0.053	0.056	0.059
indspker	-0.320	-0.294	-0.239	-0.167	-0.078	0.031	0.170	0.359	0.668
cprimary	0.161	0.157	0.145	0.128	0.108	0.083	0.052	0.010	-0.057
isecondary	0.162	0.118	0.069	0.016	-0.042	-0.106	-0.183	-0.281	-0.430
csecondary	0.220	0.163	0.100	0.034	-0.039	-0.121	-0.217	-0.339	-0.525
osecondary	0.282	0.205	0.123	0.037	-0.056	-0.160	-0.281	-0.435	-0.667
siblings	-0.018	-0.004	0.010	0.025	0.042	0.060	0.082	0.110	0.153
urban	-0.050	-0.091	-0.128	-0.164	-0.200	-0.240	-0.284	-0.340	-0.421
surban	-0.112	-0.129	-0.138	-0.144	-0.148	-0.150	-0.150	-0.147	-0.140
HFage	-0.005	-0.003	-0.002	0.000	0.002	0.005	0.008	0.011	0.017
HFcprimary	-0.033	-0.034	-0.032	-0.030	-0.026	-0.022	-0.016	-0.008	0.006
HFisecondary	0.090	0.074	0.053	0.029	0.003	-0.027	-0.064	-0.111	-0.184
HFcsecondary	0.068	0.038	0.005	-0.030	-0.068	-0.110	-0.161	-0.224	-0.321
HFosecondary	-0.074	-0.090	-0.102	-0.112	-0.120	-0.128	-0.136	-0.145	-0.155
HFincome	-0.001	0.002	0.004	0.006	0.009	0.012	0.015	0.019	0.025
AbsentFather	-0.060	-0.059	-0.054	-0.047	-0.038	-0.026	-0.012	0.007	0.038
AbsentMother	0.090	0.063	0.032	-0.001	-0.038	-0.079	-0.128	-0.191	-0.287

Figure 4.14 shows the diagnostic plots for the Poisson, Generalised-Poisson, COM-Poisson and Discrete Weibull model for the planned fertility data, confirming a good fit of the Discrete Weibull and the COM-Poisson model to this data.

The variance ratio plot in Figure 4.15 shows a good performance for the Discrete Weibull model using a linear link both on q an β.

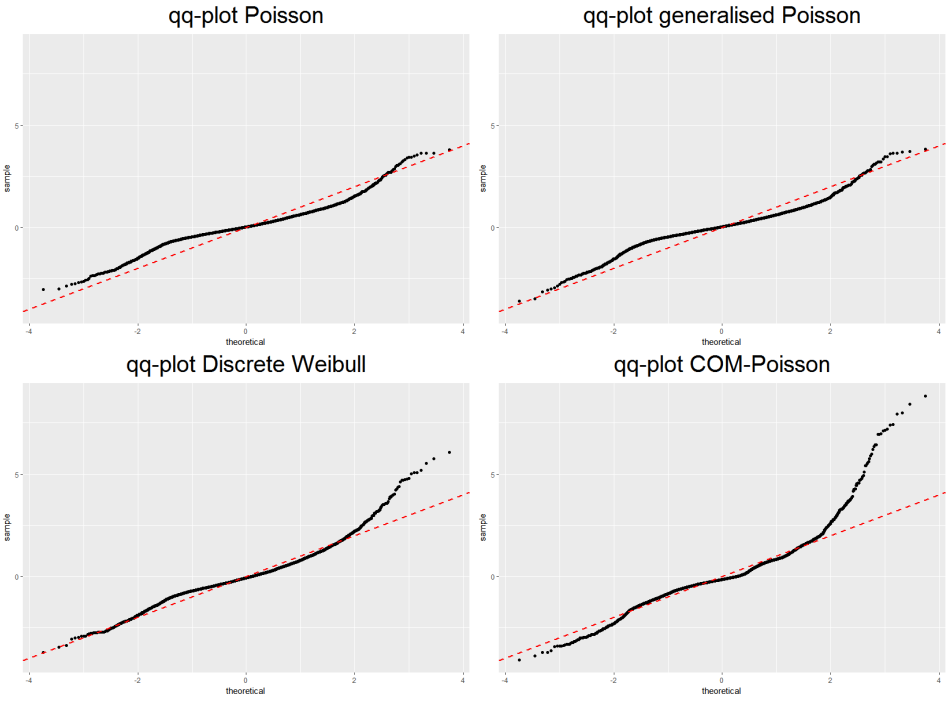


FIGURE 4.14: Diagnostic plots of the residuals for the linear models for both the regression parameters on the fertility data.

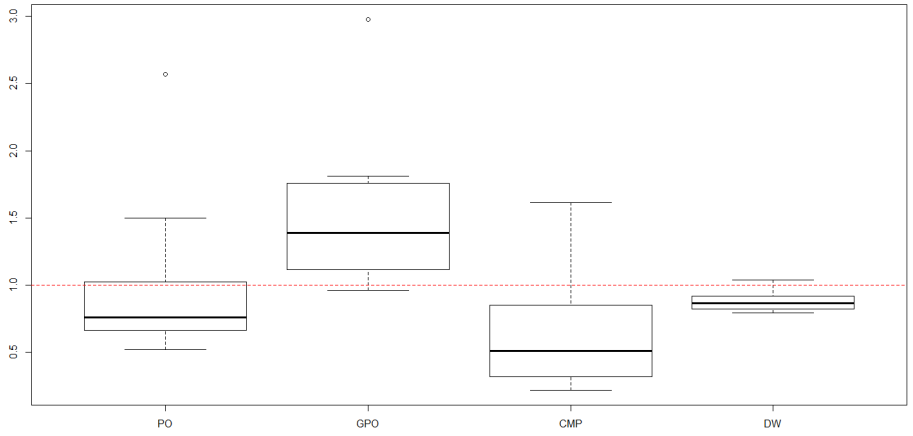


FIGURE 4.15: Variance ratio plots of the three models on the fertility data.

We further investigate these data by performing a Discrete Weibull analysis for $q(x)$ and $\beta(x)$ via the local Kernel estimator. Nevertheless, this analysis is omitted as returns similar results to the unweighted estimators, even though there is a significant improvement in terms of the AIC estimator, i.e. AIC=10024.43 for the bandwidth b=0.14. As a final analysis, we fit a non-linear Discrete Weibull model for $q(x)$ and $\beta(x)$ by including a cubic B-spline for the continuous variables SIBLINGS, HFINCOME and HFAGE. The AIC value for this non-linear model is lower than the AIC value of the linear model presented in Equation 4.16, i.e. AIC non-linear=16405.22 vs AIC linear=16415.76. There is however a large number of parameters passed into the model, i.e. 53 terms for each distributional parameters. To address this issue we maximise the L_1 penalised likelihood. The non-linear Discrete Weibull model for $q(x)$ and $\beta(x)$ with L_1 penalty has an AIC value of 16373.65. The procedure shrinks to zero a total of ten terms, i.e. three terms of the cubic B-spline of

HFINCOME, three of the cubic B-spline of HFAGE, the variable HFISECONDARY, and other three birthplace dummies. Table 4.17 reports the partial effects of this model. There does not appear to be significant changes with respect to the effects from the linear model presented in Table 4.16, and from the conclusions of this analysis presented in [69].

TABLE 4.17: Partial effects of the regressors on the dependent variable for the non-linear Jittering approach and the non-linear Discrete Weibull model with L_1 penalty for the planned fertility data.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Non-linear Jittering								
catholic	0.069	0.024	0.010	0.009	0.000	-0.006	0.014	0.037	0.060
indspker	-0.704	-0.153	-0.084	-0.055	0.013	0.071	0.120	0.215	0.605
cprimary	0.381	0.112	0.066	0.049	0.022	-0.021	-0.042	-0.110	-0.219
isecondary	0.386	0.071	0.003	-0.059	-0.107	-0.180	-0.209	-0.242	-0.426
csecondary	0.433	0.103	0.025	-0.034	-0.104	-0.192	-0.275	-0.408	-0.649
osecondary	0.419	0.089	0.018	-0.053	-0.115	-0.209	-0.290	-0.412	-0.662
cs.siblings	-0.010	0.019	0.041	0.066	0.089	0.119	0.115	0.115	0.166
urban	-0.061	-0.092	-0.134	-0.180	-0.230	-0.288	-0.340	-0.383	-0.432
surban	-0.070	-0.078	-0.096	-0.125	-0.143	-0.187	-0.205	-0.216	-0.205
cs.HFage	-0.038	-0.031	-0.023	-0.030	-0.029	-0.009	0.008	0.059	0.005
HFcprimary	-0.016	-0.010	-0.012	-0.019	-0.036	-0.047	-0.060	-0.067	-0.077
HFcsecondary	-0.031	-0.021	-0.034	-0.046	-0.074	-0.105	-0.158	-0.230	-0.324
HFosecondary	-0.083	-0.059	-0.062	-0.071	-0.090	-0.102	-0.141	-0.220	-0.325
cs.HFincome	0.172	0.196	0.179	0.308	0.334	0.387	0.498	0.781	0.936
AbsentFather	-0.043	-0.026	-0.035	-0.040	-0.056	-0.061	-0.072	-0.083	-0.091
AbsentMother	0.009	0.022	0.012	-0.010	-0.022	-0.037	-0.069	-0.161	-0.261
	Non-linear Discrete Weibull with L_1 penalty								
catholic	0.026	0.032	0.036	0.039	0.042	0.044	0.046	0.048	0.051
indspker	-0.328	-0.310	-0.257	-0.182	-0.087	0.032	0.187	0.402	0.759
cprimary	0.168	0.168	0.158	0.141	0.120	0.093	0.059	0.012	-0.064
isecondary	0.197	0.154	0.101	0.042	-0.024	-0.100	-0.191	-0.309	-0.492
csecondary	0.246	0.190	0.124	0.051	-0.030	-0.123	-0.234	-0.377	-0.597
osecondary	0.312	0.237	0.152	0.059	-0.044	-0.159	-0.297	-0.474	-0.743
cs.siblings	-0.016	0.007	0.031	0.058	0.086	0.119	0.157	0.207	0.284
urban	-0.037	-0.082	-0.123	-0.164	-0.207	-0.254	-0.308	-0.376	-0.477
surban	-0.096	-0.118	-0.132	-0.144	-0.154	-0.163	-0.172	-0.181	-0.190
cs.HFage	-0.002	-0.001	0.000	0.001	0.001	0.003	0.004	0.006	0.008
HFcprimary	-0.046	-0.045	-0.041	-0.035	-0.027	-0.018	-0.006	0.011	0.039
HFcsecondary	0.086	0.055	0.019	-0.020	-0.064	-0.114	-0.173	-0.250	-0.369
HFosecondary	-0.062	-0.083	-0.101	-0.116	-0.132	-0.147	-0.164	-0.184	-0.212
cs.HFincome	-0.008	-0.004	0.000	0.005	0.011	0.018	0.025	0.036	0.052
AbsentFather	-0.053	-0.058	-0.059	-0.058	-0.055	-0.051	-0.046	-0.037	-0.021
AbsentMother	0.044	0.028	0.010	-0.009	-0.032	-0.057	-0.088	-0.128	-0.191

4.7 Conclusions

With the analysis presented in this chapter, we contribute to the development of a flexible parametric regression model for count data. In particular, we exploit the adaptability of a Discrete Weibull distribution in modelling count data of varying dispersion, in conjunction with a generalized additive model to link its parameters to the covariates, to provide a parametric quantile regression approach for general applications with a discrete response variable. We show the applicability of this method to count data characterized by a skewed distribution making quantile regression models the preferred option for the statistical analysis of these data. Our approach can be considered as the parametric alternative to the Jittering approach of [64] and the discrete alternative to the generalized Gamma approach for continuous positive response of [74]. With respect to the Jittering approach of [64], the main difference here is that with our method the

conditional quantile function is given by a simple analytical formula, while the Jittering method employs an approximation to the unknown conditional quantile function. Another important difference with respect to the Jittering approach of [64] is that we model the conditional distribution globally via maximum likelihood, rather than via a quantile-based loss function. The advantage of these procedures for quantile-based inference is that they avoid crossing of quantiles and that they are expected to be more robust in the presence of a limited number of observations, particularly in the tails. On the other hand, parametric assumptions on the conditional distribution may limit the applicability of these approaches in situations where these are strongly violated. One last aspect of interest regards the computational time of the two approaches: our approach returns the model estimates in one step, while the quantile function via the Jittering estimator has to be computed for every τ of interest and to be averaged over a number of samples to correct the instability due to the uniform random sampling underlying the method.

Chapter 5

Conclusions

5.1 Summary

Discrete variables are those outcomes that are only allowed to assume a finite or countably infinite number of values. These variables are very common in practice, and many familiar outcomes fall into this category. Binary outcomes are widespread given that many variables and questions naturally only take two values, but also because it is often useful to construct binary variables from other types of data. Count variables recording the frequency of some events of interest are also common discrete outcomes. Because of the fundamental difference between continuous and discrete outcomes, many methods developed for continuous variables such as the popular linear least squares regression, do not apply to discrete outcomes. This thesis advanced the methodology and the application of discrete response regression models.

Policy evaluation In chapter 2, we focussed on modelling a binary response in a health policy evaluation framework. We adopted a difference-in-differences approach based on a logistic linear mixed model. Specifically, we considered multiple dependent outcomes in order to quantify the effect of the adopted pay-for-performance program while accounting for the heterogeneity of the data at the multiple nested levels. The results showed how the policy had a positive effect on the hospitals' quality in terms of those outcomes that can be more influenced by a managerial activity.

Regression models for count response In chapter 3 and in chapter 4, we focussed on modelling a count response. Typically this is done via generalised linear models [72]. The most popular approach for modelling count data is Poisson regression which assumes that the conditional distribution is Poisson with a conditional mean regressed on the covariates through the log-link function. Although Poisson regression is fundamental to the regression analysis of count data, it is often of limited use for real data, due to its property of equal mean and variance. Real data usually presents over-dispersion relative to Poisson, or the opposite case of under-dispersion. Negative Binomial regression is widely considered as the default choice for data that are over-dispersed relative to Poisson, although other options, such as the Poisson-inverse Gaussian model [111], are available. In the presence of excessive zeros, an additional component is typically added to a count distribution, such as Negative Binomial, or its truncated version, to better capture the zero generation process, leading to zero-inflated and hurdle models, respectively [18]. However, Negative Binomial regression cannot deal with data that are under-dispersed relative to Poisson. There have been some attempts to extend Poisson-based models to include also under dispersion, such as the generalised Poisson regression model [26], and the COM-Poisson regression [89]. These models are all modifications of a Poisson model and have been shown to be rather complex and computationally intensive in practice. These reasons motivated the implementation of a unified regression framework for count data via

a flexible distribution such as the Discrete Weibull. We show a number of desirable features of this distribution which are particularly appealing within a regression context: it can model both over and under-dispersed data without being restricted to either of the two; the conditional quantiles have an analytic form making the calculation of partial effects straightforward; the likelihood from a discrete Weibull model is the same as that of a continuous Weibull distribution with interval-censored data. Within the linear regression framework we have considered cases when data are grouped into clusters, or panels, or correlated groups. These models are also known in the literature as mixed-effects models. Then, the regression model has been extended to both the parameters of the distribution, and by including non-linear dependencies for both the regression parameters, and the covariates. In this way we have been able to model more accurately the full conditional distribution of Y given X, i.e. all conditional quantiles. This approach can be seen as the quantile regression alternative to available generalized linear models for counts, the parametric alternative to the Jittering approach of [64] and the discrete alternative to the generalized Gamma approach for continuous positive response of [74]. Our method has been successfully applied to simulated data and a large number of real data studies as summarised in Table 5.1, Table 5.2, and Table 5.3 for the case of over-dispersion, under-dispersion and excessive zeros respectively. These results are showing that our approach can be considered a highly competitive alternative to the current available models for count data.

TABLE 5.1: Over-dispersed data: AIC values of the Poisson, Poisson-inverse Gaussian, CMP-Poisson, Negative Binomial and Discrete Weibull models applied to different real datasets.

Data	Model	PO	PIG	CMP	NB	DW
LOS	linear (1), mixed effects	152,120.4	150,375.8	149,371.68	149,285.3	**149,210.4**
WT	non-linear (2), mixed effects	155,565.6	104,417.3	111,852.5	104,407.3	**104,362.4**
AEP	linear (2)	3,903.84	3,164.87	3,354.22	3,184.35	**3,156.27**

(1): regression model for one parameter, i.e. DW(q(x),β)
(2): regression model for both parameters, i.e. DW(q(x),β(x))

TABLE 5.2: Under-dispersed data: AIC values of the Poisson, generalised-Poisson, CMP-Poisson and Discrete Weibull models applied to different real datasets.

Data	Model	PO	GPO	CMP	DW
APGAR	linear (1), mixed effects	233,531.80	233,442.70	179,911.80	**76,761.72**
INHALER	linear (1), mixed effects	13,355.83	13,351.58	12,445.04	**12,444.26**
FERTILITY	linear (2)	18,047.07	17,118.85	16,468.79	**16,415.76**

(1): regression model for one parameter, i.e. DW(q(x),β)
(2): regression model for both parameters, i.e. DW(q(x),β(x))

TABLE 5.3: Excessive-zeros data: AIC values of the zero inflated and hurdle model formulation via Poisson, Negative Binomial and Discrete Weibull distributions applied to different real datasets.

Data	Model	ZI PO	ZI NB	ZI DW	hurdle PO	hurdle NB	hurdle DW
RWM1984	linear (1)	24,199.34	16,585.74	16,533.50	24,195.96	16,577.34	**16,528.92**
GSOEP	linear (1)	2,254.91	1,750.61	1,750.85	2,254.64	1,748.51	**1,748.42**
UPB	linear (1)	1,616.90	1,266.30	**1,265.90**	1,616.92	1,266.53	1,266.20
FISH	linear (1)	1,521.46	809.08	**802.35**	1,519.24	808.32	803.94

(1): regression model for one parameter, i.e. DW(q(x),β)

5.2 Recommendation for future research

A first consideration regards the hospital evaluation started in chapter 2 with respect to their effectiveness, and then continued in chapter 3 and chapter 4 focussing on the hospital efficiency quantified in terms of length of stay and hospital waiting times, respectively. Given the results of the three analyses it would be interesting to interpret these with respect to each hospital by ranking them and by assigning a score which reflects their performances both in terms of effectiveness and efficiency. This could offer a more representative measure of the quality of the hospitals.
Next, an interesting extension of the models for count data presented in this thesis would consider a finite mixture of Discrete Weibull distributions formed from the weighted combination of the component distributions. Moreover, it would be of interest to extend the Discrete Weibull regression models presented in this thesis to the case of a multivariate outcome, for example to offer an alternative methodological approach when facing situations such as the one presented in chapter 2. Lastly, the approach presented could additionally be investigated from a Bayesian point of view. For the linear models, this is described in [46], but mixed and non-linear models have not been developed yet in a Bayesian framework.

Appendix A

Appendix

A.1 Delta method for standard errors computation

The Delta method expands a function of a random variable around its mean, typically with a first order Taylor approximation, and then takes the variance (see Chapter 5 of [21]).

To derive the standard error of the estimators of the parameters of the Discrete Weibull with parameters $q(x)$ and β as presented in Equation 3.1, from the parametrisation of a continuous Weibull with parameters $\mu(x)$ and σ as presented in Equation 3.13 and implemented in `R` software within the `gamlss` and `survival` package, we make use of the following results.

Approximate mean and variance Let us consider a r.v. X with mean $E(X) = \mu \neq 0$ and let us assume that we want to estimate a function of μ, i.e. $g(\mu)$ with $g(\cdot)$ differentiable. Using a 1st order Taylor approximation around μ,

$$g(X) = g(\mu) + g'(\mu)(X - \mu),$$

and by using $g(X)$ as an estimator of $g(\mu)$,

$$\begin{aligned} E[g(\mu)] &= g(\mu), \\ V[g(\mu)] &= \big(g'(\mu)\big)^2 V(X). \end{aligned} \tag{A.1}$$

For example, one can consider $g(\mu) = \frac{1}{\mu}$ and a r.v. X. This leads to

$$\begin{aligned} E\left(\frac{1}{X}\right) &= \frac{1}{\mu}, \\ V\left(\frac{1}{X}\right) &= \left(\frac{1}{\mu}\right)^4 V(X). \end{aligned} \tag{A.2}$$

Moments of a ratio estimator Let us consider a r.v. X and a r.v. Y, and let us assume that we want to estimate a multivariate function $g(\mu_X, \mu_Y) = \frac{\mu_X}{\mu_Y}$, where $E(X) = \mu_X \neq 0$, $E(Y) = \mu_Y \neq 0$, $\frac{\partial}{\partial(\mu_X)} g(\cdot) = \frac{1}{\mu_Y}$, and $\frac{\partial}{\partial(\mu_Y)} g(\cdot) = -\left(\frac{\mu_X}{\mu_Y^2}\right)$. From

this

$$\begin{aligned} E\left(\frac{X}{Y}\right) &= \frac{\mu_X}{\mu_Y}, \\ V\left(\frac{X}{Y}\right) &= \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{V(X)}{\mu_X^2} + \frac{V(Y)}{\mu_Y^2} - 2\frac{\mathrm{COV}(X,Y)}{\mu_X \mu_Y}\right). \end{aligned} \tag{A.3}$$

A.1.1 `survreg` parametrisation

β coefficient: The transformation $\beta = \frac{1}{\sigma}$ leads to

$$\begin{aligned} E\left(\frac{1}{\hat{\sigma}}\right) &= \frac{1}{\sigma}, \\ V\left(\frac{1}{\hat{\sigma}}\right) &= \left(\frac{1}{\sigma}\right)^4 V(\hat{\sigma}). \end{aligned} \tag{A.4}$$

Given the nature of the link function of the scale parameter σ in the R software `survreg` parametrisation the available variance estimate is for the log estimator i.e. $V(\log\hat{\sigma})$. Thus, to derive the estimator of $V(\hat{\sigma})$ we compute

$$V(\log\hat{\sigma}) = \left(\frac{1}{\sigma}\right)^2 V(\hat{\sigma}),$$

and from this

$$V(\hat{\sigma}) = \sigma^2 V(\log\hat{\sigma}), \tag{A.5}$$

which makes possible to compute the $V\left(\frac{1}{\hat{\sigma}}\right)$ in terms of $V(\log\hat{\sigma})$ as follow

$$V\left(\frac{1}{\hat{\sigma}}\right) = \left(\frac{1}{\sigma}\right)^2 V(\log\hat{\sigma}),$$

and by replacing the unknown parameter σ with its estimator $\hat{\sigma}$.

θ parameters: The transformation $\theta = -\frac{\alpha}{\sigma}$ leads to

$$\begin{aligned} E\left(-\frac{\hat{\alpha}}{\hat{\sigma}}\right) &= -\frac{\alpha}{\sigma}, \\ V\left(-\frac{\hat{\alpha}}{\hat{\sigma}}\right) &= \left(\frac{\alpha}{\sigma}\right)^2 \left(\frac{V(-\hat{\alpha})}{\alpha^2} + \frac{V(\hat{\sigma})}{\sigma^2} - 2\frac{\mathrm{COV}(-\hat{\alpha},\hat{\sigma})}{-\alpha\sigma}\right). \end{aligned}$$

Recalling the result in Equation A.5, considering that $V(-\hat{\alpha}) = V(\hat{\alpha})$, and that $\mathrm{COV}(-\hat{\alpha},\hat{\sigma}) = E(-\hat{\alpha}\log\hat{\sigma}) - (-\alpha\log\sigma) = 0$, the variance above can be rewritten as follow

$$V\left(-\frac{\hat{\alpha}}{\hat{\sigma}}\right) = \left(\frac{\alpha}{\sigma}\right)^2 \left(\frac{V(\hat{\alpha})}{\alpha^2} + V(\log\hat{\sigma})\right),$$

and by replacing the unknown parameters α and σ with their estimators $\hat{\alpha}$ and $\hat{\sigma}$ respectively.

A.1.2 `gamlss` parametrisation

β coefficient: The transformation $\beta = \exp(\sigma)$ leads to

$$\begin{aligned} E(\exp(\hat{\sigma})) &= -\frac{\alpha}{\sigma}, \\ V(\exp(\hat{\sigma})) &= (\exp(\sigma))^2 V(\hat{\sigma}), \end{aligned} \tag{A.6}$$

where the unknown parameter σ will need to be replaced by its estimator $\hat{\sigma}$.

θ parameters: The transformation $\theta = -\alpha\exp(\sigma)$ leads to

$$E(-\hat{\alpha}\exp(\hat{\sigma})) = -\frac{\alpha}{\sigma},$$
$$V(-\hat{\alpha}\exp(\hat{\sigma})) = V\left(\frac{-\hat{\alpha}}{\frac{1}{\exp(\hat{\sigma})}}\right) = (-\alpha\exp(\sigma))^2\left(\frac{V(-\hat{\alpha})}{\alpha^2} + (\exp(\sigma))^2 V\left(\frac{1}{\exp(\hat{\sigma})}\right) - 2\frac{\mathrm{COV}(-\hat{\alpha}, \exp(\hat{\sigma}))}{-\alpha\exp(\sigma)}\right).$$

Therefore, by considering that $\mathrm{COV}(-\hat{\alpha}, \exp(\hat{\sigma})) = 0$, and using the result for $V(\exp(\hat{\sigma}))$ in Equation A.6, and the fact that

$$V\left(\frac{1}{\exp(\hat{\sigma})}\right) = \left(\frac{1}{(\exp(\sigma))^2}\right)^2 \frac{V(\exp(\hat{\sigma}))}{(\exp(\sigma))^2} = \frac{V(\hat{\sigma})}{(\exp(\sigma))^2},$$

we can express the $V(-\hat{\alpha}\exp(\hat{\sigma}))$ as follows

$$V(-\hat{\alpha}\exp(\hat{\sigma})) = (-\alpha\exp(\sigma))^2\left(\frac{V(\hat{\alpha})}{\alpha^2} + V(\hat{\sigma})\right),$$

and by replacing the unknown parameters α and σ with their estimators $\hat{\alpha}$ and $\hat{\sigma}$ respectively.

Bibliography

[1] A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

[2] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[3] Lombardy region ACT 2633. Determinazioni in ordine alla gestione del servizio socio sanitario regionale per l'esercizio 2012, 6 December 2011.

[4] Lombardy region ACT 349. Approvazione del metodo per l'individuazione dell'indice sintetico di performance per le strutture di ricovero, 23 January 2012.

[5] C. Ai and E. C. Norton. Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129, 2003.

[6] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[7] R. Alshamsan, A. Majeed, M. Ashworth, J. Car, and C. Millett. Impact of pay for performance on inequalities in health care: systematic review. *Journal of health services research and policy*, 15(3):178–184, 2010.

[8] F. J. Anscombe. Sampling theory of the Negative Binomial and logarithmic series distributions. *Biometrika*, 37(3/4):358–382, 1950.

[9] V. Apgar. A proposal for a new method of evaluation of the newborn. *Classic Papers in Critical Care*, 32(449):97, 1952.

[10] N. Atienza, J. García-Heras, J. M. Muñoz-Pichardo, and R. Villa. An application of mixture distributions in modelization of length of hospital stay. *Statistics in medicine*, 27(9):1403–1420, 2008.

[11] P. Ayyagari and D. M. Shane. Does prescription drug coverage improve mental health? Evidence from Medicare Part D. *Journal of health economics*, 41:46–58, 2015.

[12] G. P. Barbetta, G. Turati, and A. M. Zago. Behavioral differences between public and private not-for-profit hospitals in the Italian national health service. *Health economics*, 16(1):75–96, 2007.

[13] A. Barbiero. *DiscreteWeibull: Discrete Weibull Distributions (Type 1 and 3)*, 2015. R package version 1.1.

[14] P. Berta, G. Callea, G. Martini, and G. Vittadini. The effetcs of upcoding, creamskimming and readmissions on the Italian hospitals efficiency modelling: a populaiton-based investigation. *Economic modelling*, 27(4):789–890, 2010.

[15] P. Berta, G. Martini, F. Moscone, and G. Vittadini. The association between asymmetric information, hospital competition, and the quality of health care: Evidence from Italy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2016. Forhcoming.

[16] P. Berta, C. Seghieri, and G. Vittadini. Comparing health outcomes among hospitals: the experience of the Lombardy Region. *Health care management science*, 16(3):245–257, 2013.

[17] C. Bracquemond and O. Gaudoin. A survey on discrete lifetime distributions. *International Journal of Reliability, Quality and Safety Engineering*, 10(01):69–98, 2003.

[18] A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[19] J. R. Carpenter, H. Goldstein, and J. Rasbash. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):431–443, 2003.

[20] E. M. Carter and H. W. W. Potts. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC medical informatics and decision making*, 14(1):26, 2014.

[21] G. Casella and R.L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[22] C. Cashin, Y-L. Chi, P. Smith, et al. *Paying for performance in health care: implications for health system performance and accountability*. Open University Press, 2014.

[23] S. Castaldi, A. Bodina, L. Bevilacqua, E. Parravicini, M. Bertuzzi, and F. Auxilia. Payment for performance (p4p): any future in italy? *BMC public health*, 11(1):1, 2011.

[24] S. Chakraborty. Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2(1):1–30, 2015.

[25] W. S. Cleveland and S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[26] P. C. Consul and F. Famoye. Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 21(1):89–109, 1992.

[27] J. G. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844, 1971.

[28] Maladie Caisse Nationale d'Assurance. Rémunération sur objectifs de santé publique: une mobilisation des médecins et de l'assurance maladie en faveur de la qualité des soins. 2013.

[29] C. De Boor. On calculating with B-splines. *Journal of Approximation Theory*, 6(1):50–62, 1972.

[30] C. Dean, J. F. Lawless, and G. E. Willmot. A mixed Poisson–inverse-Gaussian regression model. *Canadian Journal of Statistics*, 17(2):171–181, 1989.

[31] P. Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995.

[32] J.B. Dimick and A.M. Ryan. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*, 312(22):2401–2402, 2014.

[33] P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[34] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[35] F. Eijkenaar. Pay for performance in health care: An international overview of initiatives. *Medical Care Research and Review*, 69(3):251–276, 2012.

[36] F. Eijkenaar, M. Emmert, M. Scheppach, and O. Schöffski. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health policy*, 110(2):115–130, 2013.

[37] P. HC. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

[38] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998.

[39] M. Emmert, F. Eijkenaar, H. Kemter, A. S. Esslinger, and O. Schöffski. Economic evaluation of pay-for-performance in health care: a systematic review. *The European Journal of Health Economics*, 13(6):755–767, 2012.

[40] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[41] S. J. Gange, A. Munoz, M. Saez, and J. Alonso. Use of the Beta-Binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Applied statistics*, pages 371–382, 1996.

[42] S. W. Glickman, F. S. Ou, E. R. DeLong, M. T. Roe, B. L. Lytle, J. Mulgund, J. S. Rumsfeld, W. B. Gibler, E. M. Ohman, K. A. Schulman, et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *Jama*, 297(21):2373–2380, 2007.

[43] H. Goldstein. *Multilevel statistical models*, volume 922. John Wiley and Sons, 2011.

[44] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

[45] G. K. Grunwald, S. L. Bruce, L. Jiang, M. Strand, and N. Rabinovitch. A statistical model for under-or overdispersed clustered and longitudinal count data. *Biometrical Journal*, 53(4):578–594, 2011.

[46] H. Haselimashhadi, V. Vinciotti, and K. Yu. A novel Bayesian regression model for counts with an application to health data. *Journal of Applied Statistics*, pages 1–21, 2017.

[47] T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

[48] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

[49] J. M. Hilbe. *Negative Binomial regression*. Cambridge University Press, 2011.

[50] J. M. Hilbe. *Modeling Count Data*. Cambridge University Press, 2014.

[51] M.S. Holla. Bayesian estimates of the reliability function. *Australian & New Zealand Journal of Statistics*, 8(1):32–35, 1966.

[52] J. Horowitz. A smooth binary score estimator for the binary response model. *Econometrica*, 60:505–531, 1992.

[53] J. J. Hox, M. Moerbeek, and R. Van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2010.

[54] SAS Institute Inc. Sas 9.3: Help and documentation, Cary, NC., 2012.

[55] S. Jackman. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. United States Studies Centre, University of Sydney, Sydney, New South Wales, Australia, 2017. R package V 1.5.1.

[56] A. K. Jha, K. E. Joynt, E. J. Orav, and A. M. Epstein. The long-term effect of premier pay for performance on patient outcomes. *New England Journal of Medicine*, 366(17):1606–1615, 2012.

[57] P. Karaca-Mandic, E. C. Norton, and B. Dowd. Interaction terms in nonlinear models. *Health services research*, 47(1pt1):255–274, 2012.

[58] D. Lee and T. Neocleus. Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society - Series C*, 59(5):905–920, 2010.

[59] M. Lee. Median regression for ordered discrete response. *Journal of Econometrics*, 51(1-2):59–77, 1992.

[60] Rosella Levaggi and Marcello Montefiori. Patient selection in a mixed oligopoly market for health care: the role of the soft budget constraint. *International Review of Economics*, 2013.

[61] J. Levin-Scherz, N. DeVita, and J. Timbie. Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS measures in an integrated delivery network. *Medical Care Research and Review*, 63(1 suppl):14S–28S, 2006.

[62] T. Loeys, B. Moerkerke, O. De Smet, and A. Buysse. The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180, 2012.

[63] J. Machado and M. Santos Silva. Quantiles for counts. *JASA*, 100(472):1226–1237, 2005.

[64] J. A. F. Machado and J. M. C. S. Silva. Quantiles for counts. *Journal of the American Statistical Association*, 100(472):1226–1237, 2005.

[65] C. Manski. Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 3:205–228, 1985.

[66] G. Martini, P. Berta, J. Mullahy, and G. Vittadini. The effectiveness-efficiency trade-off in health care: The case of hospitals in Lombardy, Italy. *Regional Science and Urban Economics*, 49:217–231, 2014.

[67] P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3).

[68] T. P. Minka, G. Shmeuli, J. B. Kadane, S. Borle, and P. Boatwright. Computing with the COM-Poisson distribution. *CMU Repository*, 2003.

[69] A. Miranda. Planned fertility and family background: a quantile regression for counts analysis. *Journal of Population Economics*, 21(1):67–81, 2008.

[70] J. Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

[71] T. Nagakawa and S. Osaki. The discrete Weibull distribution. *IEEE transactions on reliability*, R-24(5), 1975.

[72] J. A. Nelder and R. J. Baker. *Generalized linear models*. Wiley Online Library, 1972.

[73] J. A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74(2):221–232, 1987.

[74] A. Noufaily and M. C. Jones. Parametric quantile regression based on the generalized Gamma distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5):723–740, 2013.

[75] G. W. Oehlert. A note on the Delta method. *The American Statistician*, 46(1):27–29, 1992.

[76] L. A. Petersen, L. D. Woodard, T. Urech, C. Daw, and S. Sookanan. Does pay-for-performance improve the quality of health care? *Annals of internal medicine*, 145(4):265–272, 2006.

[77] J. Pollock. *CompGLM: Conway-Maxwell-Poisson GLM and distribution functions*, 2014. R package V 1.0.

[78] C. Propper, S. Burgess, and D. Gossage. Competition and quality: Evidence from the NHS internal market 1991–9. *The Economic Journal*, 118(525):138–170, 2008.

[79] R Core Team. R: A language and environment for statistical computing, 2014.

[80] S. Reilly, I. Olier, C. Planner, T. Doran, D. Reeves, D.M. Ashcroft, L. Gask, and E. Kontopantelis. Inequalities in physical comorbidity: a longitudinal comparative cohort study of people with severe mental illness in the uk. *BMJ Open*, 5(12), 2015.

[81] N. Rice, A. Jones, et al. Multilevel models and health economics. *Health economics*, 6(6):561–575, 1997.

[82] R. A. Rigby and D. M. Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6(1):57–65, 1996.

[83] R. A. Rigby and D. M. Stasinopoulos. *gamlss: Generalized additive models for location, scale and shape*, 2005.

[84] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

[85] F. Rousset and J. B. Ferdy. *spaMM: Testing environmental and genetic effects in the presence of spatial autocorrelation*, 2014.

[86] D. Roy. Discrete rayleigh distribution. *IEEE Transactions on Reliability*, 53(2):255–260, 2004.

[87] H. Sato, M. Ikota, A. Sugimoto, and H. Masuda. A new defect distribution metrology with a consistent discrete exponential formula and its applications. *IEEE Transactions on Semiconductor Manufacturing*, 12(4):409–418, 1999.

[88] K. F. Sellers and A. Raim. A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, 99:68–80, 2016.

[89] K. F. Sellers and G. Shmueli. A flexible regression model for count data. *Annals of Applied Statistics*, 4(2):943–961, 2010.

[90] T. Shih, L. H. Nicholas, J. R. Thumma, J. D. Birkmeyer, and J. B. Dimick. Does pay-for-performance improve surgical outcomes? An evaluation of phase 2 of the premier hospital quality incentive demonstration. *Annals of surgery*, 259(4):677, 2014.

[91] G. K. Smyth. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 47–60, 1989.

[92] T. A. B. Snijders. *Multilevel analysis*. Springer, 2011.

[93] D. M. Stasinopoulos, R. A. Rigby, et al. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007.

[94] M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani. *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press, 2017.

[95] StataCorp. Stata statistical software: Release 14. College Station, TX: StataCorp LP, 2015.

[96] C. J. Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.

[97] A. Street, C. Kobel, T. Renaud, and J. Thuilliez. How well do diagnosis-related groups explain variations in costs or length of stay among patients and across hospitals? methods for analysing routine patient data. *Health Economics*, 21(S2):6–18, 2012.

[98] M. Sutton, S. Nikolova, R. Boaden, H. Lester, R. McDonald, and M. Roland. Reduced mortality with hospital pay for performance in England. *New England Journal of Medicine*, 367(19):1821–1828, 2012.

[99] G. Szeg. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

[100] T. M. Therneau and T. Lumley. *survival: Survival Analysis*, 2017. R package V 2.41-3.

[101] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[102] P. Van Herck, D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC health services research*, 10(1):247, 2010.

[103] P. Van Herck, D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus. Systematic review: effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research*, 10(1):247, 2010.

[104] K. Västra. Assessing the impact of implementing primary care quality bonus system on follow up of patients with hypertension and type 2 diabetes based on Estonian Health Insurance Fund claims registry data in 2005-2008. 2010. *Unpublished Masters thesis, University of Tartu*, 2010.

[105] W. N. Venables and B. D. Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[106] A. P. Verbyla. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 493–508, 1993.

[107] V. Vinciotti. *DWreg: Parametric Regression for Discrete Response*, 2015. R package V 1.0.

[108] V. Vinciotti. *DWreg: Parametric Regression for Discrete Response*, 2016. R package V 2.0.

[109] J. Wang, H. Xie, and J. F. Fisher. *Multilevel models: applications using SAS*. Walter de Gruyter, 2011.

[110] R. M. Werner, J. T. Kolstad, E. A. Stuart, and D. Polsky. The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Affairs*, 30(4):690–698, 2011.

[111] G. E. Willmot. The Poisson-inverse Gaussian distribution as an alternative to the Negative Binomial. *Scandinavian Actuarial Journal*, 1987(3-4):113–127, 1987.

[112] S. Wood. *Generalized additive models: an introduction with R*. CRC press, 2006.

[113] H. Zamani and N. Ismail. Functional form for the zero-inflated generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 43(3):515–529, 2014.