

**FACULTY
OF MATHEMATICS
AND PHYSICS**
Charles University

MASTER THESIS

Marek Štěpán

Beta regression

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Šárka Hudecová, Ph.D.

Study programme: Mathematics

Study branch: Probability, Mathematical Statistics
and Econometrics

Prague 2022

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In date
Author's signature

I would like to thank my supervisor RNDr. Šárka Hudecová, Ph.D. for numerous productive consultations, professional guidance, valuable advice, and comments on this thesis.

I also want to thank my partner for the support, encouragement, and energy she has given me in completing this work.

My family has supported me throughout all my studies and I would like to express my gratitude to them as well.

Title: Beta regression

Author: Marek Štěpán

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The thesis deals with a beta regression model suitable for analysing data whose range of values is the interval $(0, 1)$. The model assumes a conditional beta distribution for the response given covariates, and its structure is similar to generalised linear models. The model is defined and its basic properties are investigated. The asymptotic distribution of the maximum likelihood estimates is provided. A possible extension to situations where the response in the data attains one of the boundary values is considered and referred to as c -inflated beta regression model. For both models, statistical inference and model diagnostics are discussed. The practical part of the thesis involves two Monte Carlo studies and two real data analyses. The first simulation study compares the performance of the global goodness-of-fit measures for link selection, while the second study explores various approaches to the analysis of the inflated beta distribution response. Alternative initial values are proposed for the cases where the algorithm did not converge. The practical usage of the model is illustrated on a model of proportions of tertiary educated people in European countries, and the proportion of household income spent on education in the Philippines.

Keywords: beta distribution, beta regression, interval data, inflated distribution

Contents

Introduction	3
1 Beta distribution	5
1.1 Definition and characteristics	5
1.1.1 Probability distribution	5
1.1.2 Relation to binomial distribution	5
1.1.3 Moments and other properties	7
1.2 Alternative parameterization	8
1.3 Extension to four parameters	10
1.4 Applications for beta distribution	10
2 Beta regression model	12
2.1 Model definition	12
2.1.1 Beta regression model definition	12
2.1.2 Connection to generalised linear models	13
2.1.3 Link function	14
2.1.4 Coefficients interpretation	14
2.1.5 Beta regression for four parameters beta distribution	16
2.2 Coefficients estimates	17
2.2.1 Score function	18
2.2.2 Fisher information matrix	19
2.2.3 Algorithm for calculation of estimates	21
2.3 Statistical inference	23
2.3.1 Regularity conditions	23
2.3.2 Asymptotic distribution and consistence	24
2.3.3 Parameters hypothesis testing	24
2.4 Model diagnostics	25
3 Extension to 0 or 1	30
3.1 Simple approaches	30
3.1.1 Omitting observations with $Y_t = c$	30
3.1.2 Adding/subtracting ε	31
3.1.3 Multiplying each Y_t	31
3.2 Inflated beta regression model	31
3.2.1 Model definition	32
3.2.2 Coefficient estimates	33
3.2.3 Statistical inference	37
3.2.4 Model diagnostics	37
3.3 Predictions	38
4 Simulations	40
4.1 Model selection and global measures	40
4.1.1 Data generation	40
4.1.2 Simulation algorithm	41
4.1.3 Results	42

4.2	Comparison of inflated beta distribution treatments	50
4.2.1	Data generation	50
4.2.2	Simulation algorithm	51
4.2.3	Results	52
5	Applications on real data	57
5.1	Proportion of tertiary educated in Europe	57
5.1.1	Data	57
5.1.2	Model fitting	58
5.1.3	Model diagnostics	59
5.2	Proportion of household income spent on education in the Philip- pines in 2015	59
5.2.1	Data	60
5.2.2	Model fitting	60
5.2.3	Model diagnostics	63
5.2.4	Comparison of treatments	63
	Conclusion	65
	Bibliography	67
	A Gamma function	69
	B Regularity conditions	70
	C Model diagnostic graphs	71
C.1	Proportion of tertiary educated in Europe	71
C.2	Proportion of household income spent on education in the Philip- pines in 2015	72

Introduction

In many real-life situations, the range of random variables can be a bounded interval. Whether it comes to a proportion of income spent, a conversion rate, a quality of wine on a given scale, or a ratio of people with a given characteristic, the beta distribution is a reasonable candidate for the distribution of such variables. Beta distribution itself has many interesting properties and one of them is that it does not belong to the natural exponential family of distributions. Therefore, generalised linear models cannot be used immediately to explore dependencies on explanatory variables. However, the beta distribution belongs to the general exponential family, and therefore most of the derivations can be used. The beta regression model was proposed by Ferrari and Cribari-Neto [2004].

In this thesis, we first take a closer look at the beta distribution and explore its properties. We investigate the relationship to the binomial distribution, moments, and provide a comprehensive description of the shapes of the probability density function. The alternative parameterization and extension to four parameters are properly defined. We provide several examples of applications where the beta distribution can occur, ranging from theoretical statistics to variables appearing in science to modern data analytics and conversion rates.

In the second chapter, the beta regression model is defined as originally proposed in Ferrari and Cribari-Neto [2004]. We discuss the connection to the generalised linear model and prove that the beta distribution does not belong to the natural exponential family, although it does belong to the (general) exponential family of distributions. We also describe and visualise the link function which can be used and seek to interpret coefficients. The main body of the chapter is dedicated to the estimation of coefficients. We calculate the score function, the Fisher information matrix, and describe an iterative algorithm for the calculation including the initial values. The derivations and calculations are described in more detail than in the original article. In the second part of the chapter, we examine statistical inference. Since the estimates are based on maximum likelihood theory, we prove the regularity conditions for beta regression and describe the inference of the maximum likelihood estimate. The last part of the chapter includes the model diagnostics described in Ferrari and Cribari-Neto [2004]. We enrich the section with an alternative pseudo R^2 and an illustration of residuals when the model is not specified well.

The third chapter describes the extension of the beta regression model for situations where the variable can also take one of the boundary values 0 or 1. Four possible solutions for such data are presented including the c -inflated beta regression model, as proposed in Ospina and Ferrari [2012]. We derive the score function and Fisher information matrix in detail and reveal their connection to the beta regression model. Statistical inference and model diagnostics are briefly described.

The second part of the thesis is practical. In Chapter 4, we perform two simulation studies. Firstly, we try to identify the best global goodness-of-fit measure, and then examine the four possible treatments of the c -inflated beta distribution. For both simulations, we describe our objectives, data generation, and simulation algorithm, and we extensively discuss the results. We also explore

the reasons why the algorithm with the initial values suggested in Ferrari and Cribari-Neto [2004] does not converge and suggest alternative initial values.

The fifth chapter of this thesis contains applications of beta regression and c -inflated beta regression to real data. We applied beta regression to a proportion of the tertiary educated population in European countries with several socioeconomic variables as covariates. FIES is a large survey in the Philippines on household income and spending habits. We explore the main drivers of the proportion of education expenditures among Filipino families.

Our own contribution to the topic of beta regression can be divided into theoretical and practical contributions. In the theoretical part of the thesis, we describe all the derivations in more detail than in the original sources. We add the interpretation of coefficients, connections to generalised linear model discussion, and the definition of the four-parameter beta distribution and how it is used in beta regression. For the inflated beta distribution, three simple approaches are summarised and compared to the c -inflated beta regression model in a simulation study. The other simulation study contradicts the original suggested global goodness-of-fit measure and shows that our alternative provides better results. Furthermore, the initial values suggested in the source articles exhibit problems with the convergence of the coefficient estimation algorithm. The alternative initial values proposed in the thesis show a significantly higher convergence rate in the simulation studies.

1. Beta distribution

In the first chapter, we recall the probability density function of a beta distribution and its characteristics with visualisations for different values of parameters. We introduce an alternative parameterization, which we will later use in the definition of the beta regression model. At the end of the chapter, we will also mention various applications where the beta distribution fits the data well.

This chapter is mainly groundwork for the beta regression model; therefore, most of the chapter is a summary of well-known facts with a focus on the interesting properties related to the beta regression. In comparison to commonly stated facts, we explore the shape of the probability density function more extensively, define the alternative parameterization, and increase the potential applications by defining the four-parameter beta distribution.

1.1 Definition and characteristics

1.1.1 Probability distribution

Definition 1. *A random variable Y follows a beta distribution if its probability density function can be written as*

$$f_Y(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, y \in (0, 1), \quad (1.1)$$

where $p > 0, q > 0$ are parameters and $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ is the gamma function. We will denote the random variable Y following the beta distribution with parameters p, q as $Y \sim B(p, q)$.

Let us show that (1.1) is a probability distribution function. For a beta function defined as $\mathcal{B}(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt$, it holds that $\mathcal{B}(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ (see Appendix A); therefore,

$$\int_0^1 f_Y(t) dt = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 t^{p-1} (1-t)^{q-1} dt = 1.$$

The cumulative distribution function of $Y \sim B(p, q)$ is

$$F_Y(y) = \frac{\mathcal{B}(y; p, q)}{\mathcal{B}(p, q)}, y \in (0, 1),$$

where $\mathcal{B}(y; p, q) = \int_0^y t^{p-1} (1-t)^{q-1} dt$ is called an incomplete beta function.

Illustrations of the distribution density for different values of parameters p, q can be found in Figure 1.1. Since $f(y; p, q) = f(1-y; q, p)$, only one of the pairs (e.g., $p = 1, q = 0.5$) of parameters is plotted; the other pair ($p = 0.5, q = 1$) would be symmetric around 0.5 – “flipped vertically”.

1.1.2 Relation to binomial distribution

There is a relationship between $B(p, q)$ and the binomial distribution. Having $X \sim \text{Bi}(n, p)$, $f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 1, 2, \dots, n$, we can see

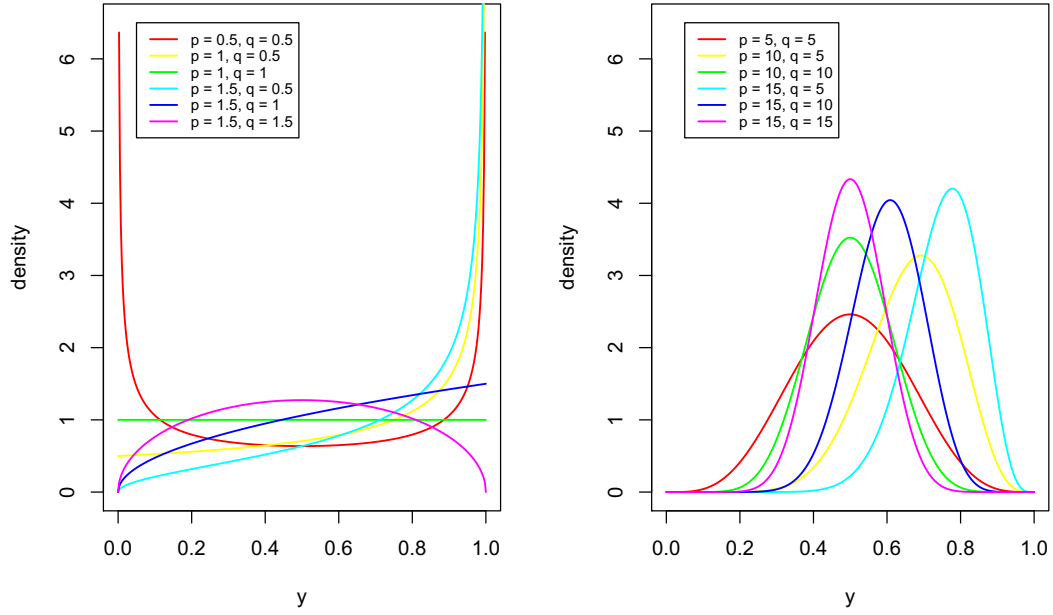


Figure 1.1: Probability density function for different choices of parameters p, q

that the distribution has a similar structure. The difference is in argument and parameters. For the binomial, the parameters are the number of trials n and the probability of success p , and the variable is the number of successes k . The beta distribution, on the other hand, models the probability and parameters are related to the number of successes and the number of failures.

Let us have $Y \sim B(p, q)$. The relation can be seen in the special case, where we choose $p, q \in \mathbb{N}, p, q > 1$ and arbitrary $y \in (0, 1)$. Knowing $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ (see Appendix A), we can see that

$$\begin{aligned}
 f_Y(y) &= \frac{(p+q-1)!}{(p-1)!(q-1)!} y^{p-1} (1-y)^{q-1} \\
 &= (p+q-1) \binom{p+q-2}{p-1} y^{p-1} (1-y)^{q-1} \\
 &= (p+q-1) f_X(p-1),
 \end{aligned}$$

where $X \sim \text{Bi}(p+q-2, y)$. This gives us an interpretation of the parameters p, q in the case $p, q = 2, 3, \dots$ – we can consider $(p-1)$ as a number of successes and $(q-1)$ as a number of failures in the $p+q-2$ attempts. The beta distribution then models (up to the multiplication constant $p+q-1$) the probability parameter given the number of successes and failures of a trial.

In Bayesian theory, the beta distribution is the conjugate prior not only for the binomial, but also for negative binomial and geometric distributions, which are all distributions derived from a sequence of independent Bernoulli trials.

1.1.3 Moments and other properties

Firstly, we are interested in the mean and variance of the beta distribution. Let us have $Y \sim B(p, q)$, it can easily be derived

$$\begin{aligned} \mathbb{E} Y &= \int_0^1 t \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} t^{p-1} (1-t)^{q-1} dt = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 t^p (1-t)^{q-1} dt \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \mathcal{B}(p+1, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{\Gamma(p+1)\Gamma(q)}{\Gamma(p+1+q)} \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{p\Gamma(p)\Gamma(q)}{(p+q)\Gamma(p+q)} = \frac{p}{p+q}, \end{aligned}$$

where we used the identity $a\Gamma(a) = \Gamma(a+1)$, one of the elemental properties of the gamma function (more about the gamma function in Appendix A).

The variance can be computed in the same way

$$\text{var}(Y) = \frac{pq}{(p+q)^2(p+q+1)}, \quad (1.2)$$

see Johnson [1995] for details.

Now, let us investigate the shape of the probability density function. We would like to describe five possible shapes: increasing, decreasing, shape with unique maximum (“hill-shape”), shape with unique minimum (“U-shape”) and a special case of the uniform distribution, where the density is constant. To do this, we need to look at the first derivative of (1.1)

$$\frac{\partial f(y)}{\partial y} = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \left(y^{p-2}(1-y)^{q-2} \right) [(p-1)(1-y) - (q-1)y],$$

because $\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} (y^{p-2}(1-y)^{q-2}) > 0$, the sign of $\frac{\partial f(y)}{\partial y}$ depends only on

$$g(y) = (p-1)(1-y) - (q-1)y = (p-1) - y(p+q-2).$$

If $p+q = 2$, we have $g(y) = (p-1)$ and for all $y \in (0, 1)$ it holds that

- $\frac{\partial f(y)}{\partial y} = 0$ (density is constant) if $p = 1$, which further implies $q = 1$ (special case – uniform distribution),
- $\frac{\partial f(y)}{\partial y} > 0$ (density is increasing) if $p > 1$ which implies that $q < 1$,
- $\frac{\partial f(y)}{\partial y} < 0$ (density is decreasing) if $p < 1$ which implies that $q > 1$.

When $p+q \neq 2$, $f_Y(y)$ has a local extremum at the point $\frac{p-1}{p+q-2}$, because the sign of $g(y)$ (and therefore $\frac{\partial f(y)}{\partial y}$) is different for $y < \frac{p-1}{p+q-2}$ and $y > \frac{p-1}{p+q-2}$. Denote the local extreme point $c = \frac{p-1}{p+q-2}$.

If $p+q > 2$, then we have

- $\frac{\partial f(y)}{\partial y} > 0$ (density is increasing) for $y < c$,
- $\frac{\partial f(y)}{\partial y} < 0$ (density is decreasing) for $y > c$.

If $p+q < 2$, then we have

- $\frac{\partial f(y)}{\partial y} < 0$ (density is decreasing) for $y < c$,
- $\frac{\partial f(y)}{\partial y} > 0$ (density is increasing) for $y > c$.

Putting those results together with the position of the extremum ($c \leq 0$, $c \in (0, 1)$ or $c \geq 1$) the following can be said about the shape of the probability density function of the beta distribution

- increasing for $p > 1, q < 1$,
- decreasing for $p < 1, q > 1$,
- “hill-shaped” with maximum (mode of the distribution) at $c = \frac{p-1}{p+q-2}$ for $p \geq 1, q \geq 1, p \neq q$,
- “U-shaped” with minimum (anti-mode of the distribution) at $c = \frac{p-1}{p+q-2}$ for $p \leq 1, q \leq 1, p \neq q$,
- constant for $p = 1, q = 1$ (the uniform distribution),

all of those shapes can be seen in Figure 1.1.

The limits around the boundary points $\{0, 1\}$ can be calculated directly from the formula (1.1) of the probability density function using the identity $a\Gamma(a) = \Gamma(a+1)$ and the fact that $\Gamma(1) = 1$ (see Appendix A)

$$\lim_{y \rightarrow 0^+} f_Y(y) = \begin{cases} 0, & \text{if } p > 1, \\ \frac{\Gamma(q+1)}{\Gamma(q)} = q, & \text{if } p = 1, \\ \infty, & \text{if } p < 1. \end{cases}$$

$$\lim_{y \rightarrow 1^-} f_Y(y) = \begin{cases} 0, & \text{if } q > 1, \\ \frac{\Gamma(p+1)}{\Gamma(p)} = p, & \text{if } q = 1, \\ \infty, & \text{if } q < 1. \end{cases}$$

In the special case $p = q$ (or equivalently $\mu = \mathbf{E} Y = \frac{1}{2}$), the density is symmetric around $\frac{1}{2}$. Symmetry $f_Y(y) = f_Y(1-y)$ can be seen directly from the probability density function.

1.2 Alternative parameterization

In a beta regression model, which will be introduced in Chapter 2, a different possible parameterization of the beta distribution is used. Since we want to model a conditional mean, the first parameter will be the mean $\mu = \mathbf{E} Y$. The second parameter ϕ is called the dispersion parameter. The alternative parameters μ, ϕ are defined by the parameters p, q as

$$\mu = \frac{p}{p+q}, \tag{1.3}$$

$$\phi = p+q. \tag{1.4}$$

We can see that $\mu \in (0, 1)$, $\phi > 0$. Parameters p, q can be derived back as

$$\begin{aligned}\mu &= \frac{p}{p+q} = \frac{p}{\phi} \\ p &= \mu\phi \\ q &= \phi - p = \phi - \mu\phi = (1 - \mu)\phi,\end{aligned}$$

where we used (1.4) for the derivation of q .

The probability density function under this alternative parameterization can be written as

$$f_Y(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad y \in (0, 1). \quad (1.5)$$

The cumulative distribution function can be obtained in a similar way by substituting the parameters p, q by $\mu\phi, (1-\mu)\phi$.

Illustrations of the distribution density for different values of parameters μ, ϕ can be found in Figure 1.2. Similarly to the parameterization of p, q , we have symmetry $f(y; \mu, \phi) = f(1-y; 1-\mu, \phi)$, and only $\mu \geq 0.5$ cases are plotted. For $\mu < 0.5$ the PDF would be symmetrical around 0.5.

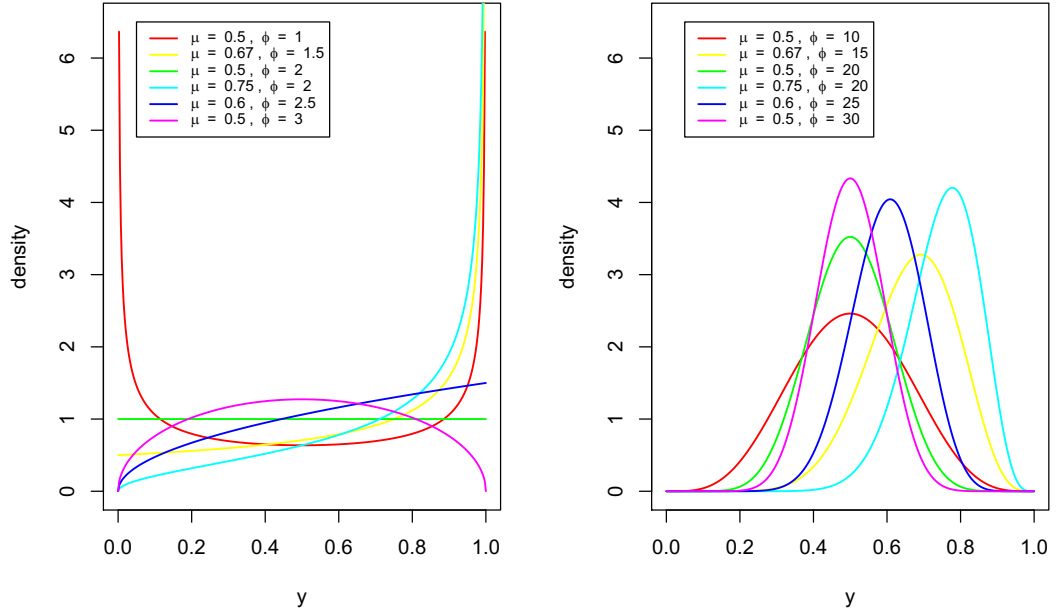


Figure 1.2: Probability density function under alternative parameterization for different choices of parameters μ, ϕ ; the figure corresponds to Figure 1.1 to compare different parameterizations for the same density shapes

The variance under the alternative parameterization is derived by substituting p, q in (1.2) by μ, ϕ

$$\begin{aligned}\text{var}(Y) &= \frac{pq}{(p+q)^2(1+p+q)} = \frac{p}{p+q} \frac{q}{p+q} \frac{1}{1+p+q} = \mu(1-\mu) \frac{1}{1+\phi} \\ &= \frac{\mu(1-\mu)}{1+\phi}.\end{aligned} \quad (1.6)$$

Now we can see why the name dispersion parameter for ϕ makes sense: for fixed μ , the variance is smaller for larger ϕ and vice versa. For the alternative parameterization, the beta distribution is symmetric when $\mu = \frac{1}{2}$, which can be easily seen by the definition of μ and the symmetry condition for standard p, q parameterization. Furthermore, the shape of the probability density function can be categorised as previously stated by substituting p, q for μ, ϕ .

1.3 Extension to four parameters

The beta distribution is originally defined in the interval $(0, 1)$, but there is a straightforward extension to an arbitrary interval (a, b) .

Definition 2. Let $Y \sim B(p, q)$ and $a, b \in \mathbb{R}, a < b$ be parameters. Then we say that the random variable $\tilde{Y} = a + (b - a)Y$ follows the four-parameter beta distribution with parameters p, q, a, b and write $\tilde{Y} \sim B(p, q, a, b)$.

The probability density function of $\tilde{Y} \sim B(p, q, a, b)$ can be derived using the inverse relationship $Y = \frac{\tilde{Y} - a}{b - a}$ as follows

$$\begin{aligned} f_{\tilde{Y}}(\tilde{y}) &= \frac{1}{b - a} f_Y\left(\frac{\tilde{y} - a}{b - a}\right) \\ &= \frac{1}{b - a} \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} \left(\frac{\tilde{y} - a}{b - a}\right)^{p-1} \left(1 - \frac{\tilde{y} - a}{b - a}\right)^{q-1}, \\ &= \frac{1}{b - a} \frac{1}{(b - a)^{p-1}} \frac{1}{(b - a)^{q-1}} \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} (\tilde{y} - a)^{p-1} (b - \tilde{y})^{q-1}, \\ &= \frac{1}{(b - a)^{p+q-1}} \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} (\tilde{y} - a)^{p-1} (b - \tilde{y})^{q-1}, \end{aligned}$$

for $\tilde{y} \in (a, b), p, q > 0$.

This extension significantly increases the variety of applications. With a four-parameter beta distribution, we can model not only variables from range $(0, 1)$ but from any arbitrary finite interval.

Parameters a, b can be known (given by experimental conditions) or unknown and subject to statistical inference. In this work, we will consider only the first case where a, b are given beforehand – more about the other case can be found in Johnson [1995]. The use of a beta regression model for this extension is described in Section 2.1.5.

1.4 Applications for beta distribution

The beta distribution is well suited to model the behaviour of random variables with a domain limited to the interval $(0, 1)$ and the straightforward extension to the four-parameter beta distribution widens the range of applications to variables with finite interval domains.

There are four types of beta distribution density (illustrated in Figure 1.1 or Figure 1.2) – with a single mode, with a single anti-mode, monotonous, and constant in the interval $(0, 1)$. This gives us a great compromise between simplicity and applicability for most of the finite-interval-bounded variables.

In statistical theory, the beta distribution arises in several fields. An example is order statistics for a random sample from a continuous uniform distribution on $(0, 1)$. The distribution for the k -th smallest observation of n random variables has a beta distribution $B(k, n + 1 - k)$. A proof of this statement can be found, for example, in David [2003].

Another important application can be found in situations where statements are not true/false with 100% certainty. These are the cases where true or false depend on the reliability of the source. The underlying theory is called subjective logic, and the beta distribution is used to model this situation, more in Jøsang [2016].

More applications of beta distribution can be found in project management, wavelet analysis, and modelling distribution of probabilities. A large group of real-life applications is proportions, which arise in case our data are aggregated by groups and we do not have a record for each individual.

The beta distribution is well suited for modelling in science (a proportion of particles filtered, how efficient is a chemical decomposition or a success rate of distillation) as well as in socioeconomics (a proportion of families below the poverty line, college graduation chance for children, or food waste rate in a specific region). In modern data analytics, the beta distribution can be used to model conversion rates (e-commerce purchase conversion, newsletter reading rate, or advertisement conversion to visit a website).

In addition to the applications of the two-parameter beta distribution, the four-parameter beta distribution gives us a model for an arbitrary finite-interval bounded variable. There are many examples, from income distribution (often modelled as a bounded random variable), fulfilment of a warehouse, to RGB-decomposed pixel in a picture.

In Chapter 3 we examine the situation where the random variable is not only in $(0, 1)$, but can also be equal to 0 or 1. This case arises in many applications, such as the mortality rate of car accidents, attendance at an event, the percentage of students who pass a college year, or a proportion of education expenditures.

In general, we can say that for random variables with finite interval domain, modelling randomness using the beta distribution can be a good idea, if its distribution has one centre of mass (mode for $p, q \geq 1, p \neq q$) or the mass is concentrated around both boundaries (anti-mode for $p, q \leq 1, p \neq q$) or if we can assume monotonous or constant density; illustrations are shown in Figure 1.1.

2. Beta regression model

As we mentioned in the previous section, the beta distribution can be applied to model various random variables. Many times those variables depend on other factors, and below we describe the classic regression-style dependency, where the conditional mean of the response given the covariates is examined. This is a case of many applications mentioned; in science, we often have conditions of the experiment as covariates, and in socioeconomics, the response often depends on age, a district, or a density of a population. For all of those cases, a beta regression may be an appropriate model.

In this section, we will define the beta regression model, estimate its coefficients, propose diagnostic methods, and derive confidence intervals and tests based on maximum likelihood theory. The main source for this chapter is the article Ferrari and Cribari-Neto [2004], where the model is introduced, the score function and the Fisher information matrix are derived, statistical inference is briefly described, and diagnostic tools are proposed. In this work, we formally define the beta regression model using stochastic covariates (in contrast to Ferrari and Cribari-Neto [2004]), discuss its connection to generalised linear models and interpret the coefficients. The section describing the estimation of coefficients is enriched with a detailed derivation, and an algorithm for the calculation is described in detail. In the section Statistical inference, regularity conditions are verified and, as in Ferrari and Cribari-Neto [2004], asymptotic results are described including the test and confidence interval for regression parameters. In diagnostics, we state the most important tools of the article and enhance it with a particular example of usage.

2.1 Model definition

2.1.1 Beta regression model definition

Definition 3. *Data $(Y_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$ satisfy the beta regression model if*

1. *$(Y_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$ are independent identically distributed random vectors and $P(\text{rank}(\mathbb{X}) = k) = 1$, where $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$.*
2. *The conditional distribution $Y_t | \mathbf{X}_t \sim B(\mu_t, \phi), t = 1, \dots, n$, where $\mu_t, t \in \{1, 2, \dots, n\}$, and ϕ are unknown parameters.*
3. *For all $t = 1, \dots, n$ μ_t depends on $\mathbf{X}_t$ through a linear predictor $\eta_t = \mathbf{X}_t^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^\top$ are unknown regression parameters.*
4. *There exists strictly monotonous, twice differentiable function $g(\cdot)$ mapping $(0, 1)$ to $\mathbb{R}$ which links conditional means μ_t and linear predictors η_t*

$$g(\mu_t) = \eta_t = \sum_{i=1}^k X_{ti} \beta_i = \mathbf{X}_t^\top \boldsymbol{\beta}.$$

The first covariate $(X_{t1}, t \in \{1, 2, \dots, n\})$ is often set to 1 as the intercept of the model. Note that we use the notation starting from 1, so β_1 stands for the

intercept coefficient, and the total number of coefficients and the length of the vector $\mathbf{X}$ is k .

We can directly see that our model will be heteroskedastic, which means that Y_t has a non-constant conditional variance. The conditional variance of our response depends on the mean and, therefore, also on the covariates.

$$\text{var}(Y_t|\mathbf{X}_t) = \frac{\mu_t(1 - \mu_t)}{1 + \phi} = \frac{g^{-1}(\mathbf{X}_t^\top \boldsymbol{\beta})(1 - g^{-1}(\mathbf{X}_t^\top \boldsymbol{\beta}))}{1 + \phi}. \quad (2.1)$$

Remark. Let us denote here that we define the model using random covariates. In contrast to this approach, some authors (e.g., Ferrari and Cribari-Neto [2004]) define the beta regression for fixed covariates. The results presented in this thesis also hold for fixed covariates, but the derivation and some assumptions would be slightly different.

2.1.2 Connection to generalised linear models

Looking at Definition 3, there is a similarity between the definition of the beta regression model and generalised linear models (the definition and the corresponding theory can be found, for example, in McCullagh and Nelder [1989]). The only difference is the assumption of the beta distribution and therefore absence of the canonical parameter – for the generalised linear model, the conditional distribution is assumed to belong to a family of natural exponential distributions with the canonical parameter θ_i depending on $\mathbf{X}_t$ and $\boldsymbol{\beta}$, and the dispersion parameter φ

$$f(y; \theta_t, \varphi) = \exp \left\{ \frac{y\theta_t - b(\theta_t)}{\varphi} + c(y, \varphi) \right\}.$$

The beta regression distribution is not a member of the natural exponential family, because the exponential type requires a linear element y in the exponent ($y\theta_t$), which is not possible for the beta distribution, because when rewriting the density of the beta distribution in exponential form, y appears only transformed as $\log(y)$ or $\log(1 - y)$

$$\begin{aligned} f_Y(y) &= \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1} \\ &= \exp \left\{ (\mu\phi - 1) \log(y) + [(1-\mu)\phi - 1] \log(1-y) \right. \\ &\quad \left. + \log(\Gamma(\phi)) - \log(\Gamma(\mu\phi)) - \log(\Gamma((1-\mu)\phi)) \right\}, \end{aligned} \quad (2.2)$$

therefore the beta distribution does not belong to the natural exponential distribution family.

However, the beta distribution belongs to the (general) exponential family of distributions, which can be defined as in Lehmann and Casella [1998] as distributions with densities of the form

$$f(y; \boldsymbol{\theta}) = \exp \left[\sum_{i=1}^s \eta_i(\boldsymbol{\theta}) T_i(y) - B(\boldsymbol{\theta}) \right] h(y), \quad (2.3)$$

where $\eta_i(\cdot)$ and $B(\cdot)$ are real-valued functions, $T_i(y)$ are real-valued statistics. Having $\boldsymbol{\theta} = (\mu, \phi)^\top$ we can derive using (2.2)

$$\begin{aligned}\eta_1(\boldsymbol{\theta}) &= \mu\phi, \\ \eta_2(\boldsymbol{\theta}) &= (1 - \mu)\phi, \\ T_1(y) &= \log y, \\ T_2(y) &= \log(1 - y), \\ B(\boldsymbol{\theta}) &= \log \left[\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1 - \mu)\phi)} \right], \\ h(y) &= \frac{1}{y(1 - y)}.\end{aligned}$$

This gives us motivation to explore the theory of the beta regression model (given in Definition 3) similarly to the theory of generalised linear models.

2.1.3 Link function

The function $g(\mu)$ is called a *link function* as for generalised linear models. There are several options for the link function:

- logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$,
- probit link $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution,
- complementary log-log link $g(\mu) = \log(-\log(1 - \mu))$,
- log-log link $g(\mu) = -\log(-\log(\mu))$,
- Cauchy link $g(\mu) = \tan(\pi(\mu - 0.5))$.

In Figure 2.1 there are illustrations of the suggested link functions. Assuming an intercept in the model, we shift the complementary log-log link and log-log link to 0 at $\mu_t = \frac{1}{2}$ for a clearer comparison.

The most commonly used link function is the logit link. In many applications, the selection of the link function does not substantially change the measure of fit (e.g. pseudo R^2 introduced later), since the shapes of the suggested links are similar around $\mu_t = \frac{1}{2}$. This was also observed in the simulation study in Section 4.1. Differences in link functions are more distinct around the boundary values, as can be seen in Figure 2.2.

A diagnostic plot suitable to verify the choice of the link function is described in Section 2.4.

2.1.4 Coefficients interpretation

In this section, the interpretation of the regression coefficients is discussed to see how the ratios or the coefficients themselves can be derived from different perspectives.

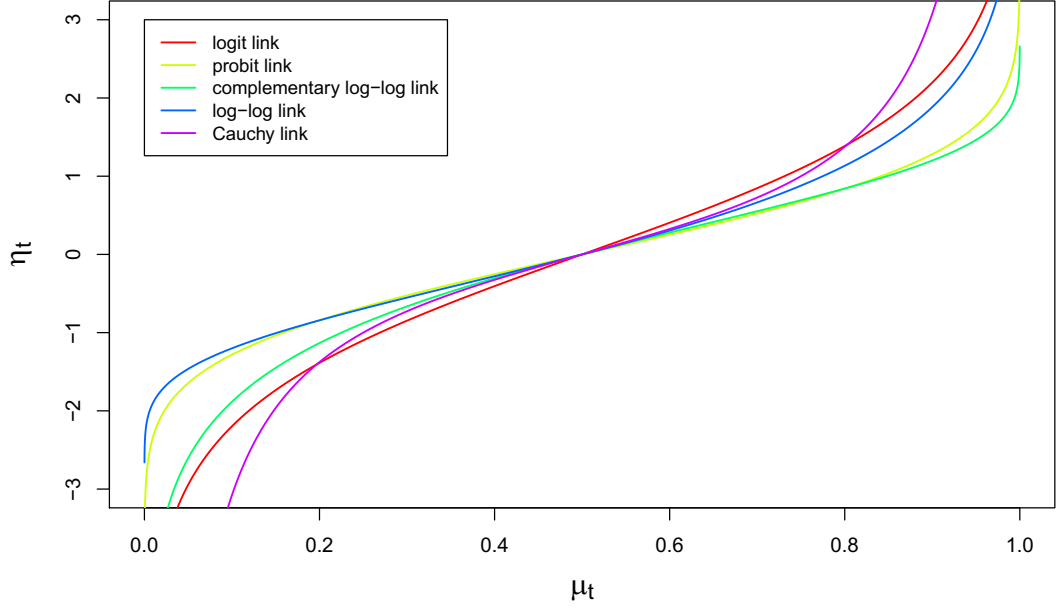


Figure 2.1: Illustrations of the suggested link functions

The first general interpretation is derived for the ratio of the coefficients $\beta_i, \beta_j, i, j \in \{1, \dots, k\}, i \neq j$. Since

$$\frac{\partial \mu_t}{\partial X_{ti}} = \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial X_{ti}} = \frac{1}{g'(\mu_t)} \frac{\partial \mathbf{X}_t^\top \boldsymbol{\beta}}{\partial X_{ti}} = \frac{1}{g'(\mu_t)} \beta_i,$$

the ratio of the regression coefficients $\frac{\beta_i}{\beta_j}$ is equal to the ratio of the marginal effect of the regressors

$$\frac{\beta_i}{\beta_j} = \frac{\frac{1}{g'(\mu_t)} \beta_i}{\frac{1}{g'(\mu_t)} \beta_j} = \frac{\frac{\partial \mu_t}{\partial X_{ti}}}{\frac{\partial \mu_t}{\partial X_{tj}}}.$$

The identity can be interpreted as the ratio of effects caused by marginal changes in the regressors X_i, X_j . In this interpretation, care must be taken with regard to the units of the variables.

The logit link is probably the most widely used link function due to the straightforward interpretability of the model coefficients. Similarly to logistic regression, we can interpret the coefficient as the logarithm of the ratio of $\frac{\mu}{1-\mu}$ (sometimes called the odds ratio as in logistic regression). We know that

$$\frac{\mu_t}{1 - \mu_t} = e^{\eta_t} = e^{\mathbf{X}_t^\top \boldsymbol{\beta}},$$

so let us have two vectors of the realisation of covariates $\mathbf{x}_0 = (x_{01}, \dots, x_{0k})$, $\mathbf{x}_1 = (x_{01}, \dots, x_{0i} + 1, \dots, x_{0k})$, so that the vector $\mathbf{x}_1$ is a vector with element i increased by 1 and the other elements remain unchanged. Denote $E[Y_t | \mathbf{X}_t = \mathbf{x}_0] = \mu_0$ and $E[Y_t | \mathbf{X}_t = \mathbf{x}_1] = \mu_1$, then

$$\frac{\frac{\mu_1}{1-\mu_1}}{\frac{\mu_0}{1-\mu_0}} = \frac{e^{\mathbf{x}_1^\top \boldsymbol{\beta}}}{e^{\mathbf{x}_0^\top \boldsymbol{\beta}}} = e^{\beta_i}.$$

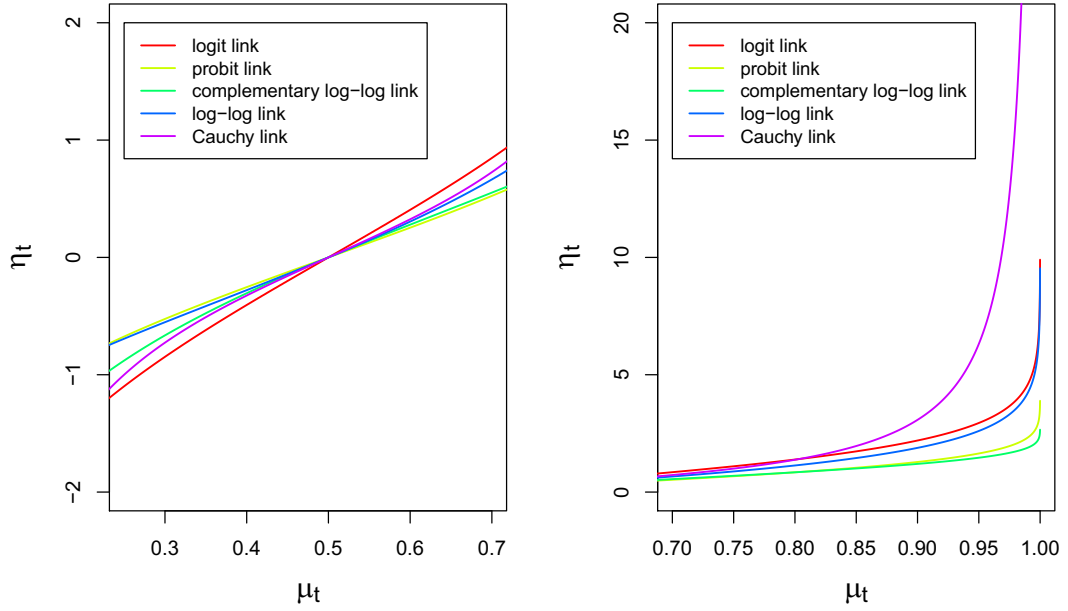


Figure 2.2: Illustrations of the similarity of the suggested link functions around $\mu_t = \frac{1}{2}$ and the distinction for $\mu_t \in (0.8, 1)$

We can see that a unit change in X_{ti} causes the logarithm of the ratio of transformed conditional means $\frac{\mu}{1-\mu}$ to increase by β_i . Note that $\frac{\frac{\mu_1}{1-\mu_1}}{\frac{\mu_0}{1-\mu_0}}$ is not the proper odds ratio as in logistic regression, because μ denotes the mean instead of the probability of an event. However, we can still easily calculate the effect of a particular coefficient on the transformed response ratio. The interpretation of the ratio can be seen if we model proportions. Then μ denotes the mean proportion of the characteristic among the group and $1 - \mu$ denotes the proportion of the population without that characteristic. The ratio $\frac{\mu}{1-\mu}$ can then be interpreted as a ratio of those proportions, and e^{β_i} measures how this ratio increases if X_{ti} increases by 1 unit.

The intercept coefficient β_1 does not have a special interpretation rather than the usual interpretation for a situation where all other covariates are equal to 0; this is the situation where we only have the intercept and $\hat{\mu}_t = \hat{\mu}$ estimates the mean of Y_t transformed by the link function $g(\cdot)$.

2.1.5 Beta regression for four parameters beta distribution

In Section 1.3 we defined the four-parameter beta distribution. In this section, we briefly discuss how the beta regression model can be used for a response following the four-parameter beta distribution.

Let us consider the case where we have data $(\tilde{Y}_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$ that satisfy the beta regression model (Definition 3) with $\mathcal{Z}_*$ replaced by $\mathcal{Z}_*^*$. $\tilde{Y}_t | \mathbf{X}_t \sim B(\mu_t, \phi, a, b), t = 1, \dots, n$. Then we can do a simple transformation using the fact

that $\tilde{Y} = a + (b - a)Y$, where $Y \sim B(\mu_t, \phi)$ and apply the model for the beta regression model for the data $(Y_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$, where $Y_t = \frac{\tilde{Y}_t - a}{b - a}$. Since

$$\mathbb{E}[\tilde{Y}_t | \mathbf{X}_t] = a + (b - a) \mathbb{E}[Y_t | \mathbf{X}_t],$$

we can easily calculate the predictions for the original data. Using the same transformation, the confidence interval for the conditional mean can also be transformed.

Note that a special case of the four-parameter beta distribution is variables with the range of values $(0, b)$. This situation represents the fact that the original data have a different unit of measurement that follows the same scale. Such cases can be seen for larger values, where we often transform the variable to represent thousands, millions, etc. Also, data where proportions (percentage) are recorded on a scale 0 to 100 belong to this special case with $a = 0, b = 100$.

2.2 Coefficients estimates

In this section, we derive the maximum likelihood estimates for the model parameters β, ϕ , calculate the score function, Fisher's information matrix, and describe the Fisher scoring algorithm for the calculation of the estimates.

Having $\theta = (\beta^\top, \phi)^\top$, $\Theta = \mathbb{R}^k \times (0, \infty)$, the maximum likelihood estimate of $\hat{\theta} = (\hat{\beta}^\top, \hat{\phi})^\top$ is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}^*(\theta),$$

where $\mathcal{L}^*(\theta)$ is the likelihood function of our data.

Let $(Y_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$ be data that satisfy the definition of the beta regression model (Definition 3) with true parameters β^0, ϕ^0 . The full likelihood of the data is the following

$$\begin{aligned} \mathcal{L}^*(\beta, \phi) &= \prod_{t=1}^n f_{Y, \mathbf{X}}(Y_t, \mathbf{X}_t) = \prod_{t=1}^n f_{Y|\mathbf{X}}(Y_t, \mathbf{X}_t) f_{\mathbf{X}}(\mathbf{X}_t) \\ &= \left(\prod_{t=1}^n f_{Y|\mathbf{X}}(Y_t, \mathbf{X}_t) \right) \left(\prod_{t=1}^n f_{\mathbf{X}}(\mathbf{X}_t) \right), \end{aligned}$$

where $f_{Y, \mathbf{X}}(Y_t, \mathbf{X}_t)$ is the joint density of $(Y_t, \mathbf{X}_t^\top)^\top$, $f_{Y|\mathbf{X}}(Y_t, \mathbf{X}_t)$ is the conditional density of $Y_t | \mathbf{X}_t \sim B(\mu_t, \phi)$, and $f_{\mathbf{X}}(\mathbf{X}_t)$ is the marginal density of $\mathbf{X}_t$. Since we want to maximise $\mathcal{L}^*(\beta, \phi)$ with respect to β, ϕ , it suffices to maximise

$$\mathcal{L}(\beta, \phi) = \prod_{t=1}^n f_{Y|\mathbf{X}}(Y_t, \mathbf{X}_t),$$

because $f_{\mathbf{X}}(\mathbf{X}_t)$ does not depend on β and ϕ . The logarithm is a monotonous function; therefore, it is sufficient to maximise the log-likelihood

$$\ell(\beta, \phi) = \sum_{t=1}^n \log \left(f_{Y|\mathbf{X}}(Y_t, \mathbf{X}_t) \right).$$

Denote $\ell_t(\boldsymbol{\beta}, \phi) = \log(f_{Y|X}(Y_t, \mathbf{X}_t))$ and derive

$$\begin{aligned}
\ell_t(\boldsymbol{\beta}, \phi) &= \log \left(\frac{\Gamma(\phi)}{\Gamma(\mu_t \phi) \Gamma((1 - \mu_t) \phi)} Y_t^{\mu_t \phi - 1} (1 - Y_t)^{(1 - \mu_t) \phi - 1} \right) \\
&= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1 - \mu_t) \phi) + (\mu_t \phi - 1) \log Y_t \\
&\quad + [(1 - \mu_t) \phi - 1] \log (1 - Y_t) \\
&= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1 - \mu_t) \phi) + \mu_t \phi \log Y_t - \log Y_t \\
&\quad + \phi \log(1 - Y_t) - \mu_t \phi \log(1 - Y_t) - \log(1 - Y_t), \tag{2.4}
\end{aligned}$$

so

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \phi) &= \sum_{t=1}^n \ell_t(\boldsymbol{\beta}, \phi) = \sum_{t=1}^n \left\{ \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1 - \mu_t) \phi) \right. \\
&\quad \left. + \mu_t \phi \log Y_t - \log Y_t + \phi \log(1 - Y_t) - \mu_t \phi \log(1 - Y_t) - \log(1 - Y_t) \right\}. \tag{2.5}
\end{aligned}$$

The maximum likelihood estimator for the coefficients in the beta regression model is then defined as

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\phi})^\top = \underset{(\boldsymbol{\beta}^\top, \phi)^\top \in \mathbb{R}^k \times (0, \infty)}{\operatorname{argmax}} \ell(\boldsymbol{\beta}, \phi).$$

2.2.1 Score function

In the following two sections we derive the score vector and the Fisher information matrix, because the maximum likelihood estimate can be found as a solution to the equations

$$\mathbf{U}(\boldsymbol{\beta}, \phi) = \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\beta}^\top, \phi)^\top} = \mathbf{0},$$

where $\mathbf{U}(\boldsymbol{\beta}, \phi)$ is called a score function.

Let us denote the two elements of the score function corresponding to $\boldsymbol{\beta}$, ϕ as

$$\mathbf{U}(\boldsymbol{\beta}, \phi) = (\mathbf{U}_\beta(\boldsymbol{\beta}, \phi)^\top, U_\phi(\boldsymbol{\beta}, \phi))^\top,$$

where $\mathbf{U}_\beta(\boldsymbol{\beta}, \phi) = (U_{\beta_1}(\boldsymbol{\beta}, \phi), \dots, U_{\beta_k}(\boldsymbol{\beta}, \phi))^\top$ and calculate for $i \in \{1, \dots, k\}$

$$U_{\beta_i}(\boldsymbol{\beta}, \phi) = \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_i} = \sum_{t=1}^n \frac{\partial \ell_t(\boldsymbol{\beta}, \phi)}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_i}.$$

See that $\frac{\partial \mu_t}{\partial \eta_t} = \frac{1}{g'(\mu_t)}$ and $\frac{\partial \eta_t}{\partial \beta_i} = X_{ti}$ and for the rest we use the digamma function ψ (see Appendix A) and calculate the derivative

$$\begin{aligned}
\frac{\partial \ell_t(\boldsymbol{\beta}, \phi)}{\partial \mu_t} &= -\psi(\mu_t \phi) \phi - \psi((1 - \mu_t) \phi) (-\phi) + \phi \log Y_t - \phi \log(1 - Y_t) \\
&= \phi \left[\log \frac{Y_t}{1 - Y_t} - (\psi(\mu_t \phi) - \psi((1 - \mu_t) \phi)) \right].
\end{aligned}$$

For simplification of further calculations, we denote

$$Y_t^* = \log \frac{Y_t}{1 - Y_t}, \quad \mu_t^* = \psi(\mu_t \phi) - \psi((1 - \mu_t) \phi),$$

and now we can complete the first element of the score function

$$U_{\beta_i}(\boldsymbol{\beta}, \phi) = \phi \sum_{t=1}^n (Y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} X_{ti}.$$

If we take $\mathbf{Y}^* = (Y_1^*, \dots, Y_n^*)^\top$, $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_n^*)^\top$, $\mathbb{T} = \text{diag}\left(\frac{1}{g'(\mu_1)}, \dots, \frac{1}{g'(\mu_n)}\right)$ and $\mathbb{X} \in \mathbb{R}^{n \times k}$ as the regression matrix (with $\mathbf{X}_t^\top, t = 1, \dots, n$ in the rows), we have

$$\mathbf{U}_\beta(\boldsymbol{\beta}, \phi) = \phi \mathbb{X}^\top \mathbb{T}(\mathbf{Y}^* - \boldsymbol{\mu}^*). \quad (2.6)$$

For $U_\phi(\boldsymbol{\beta}, \phi)$ we derive

$$\begin{aligned} U_\phi(\boldsymbol{\beta}, \phi) &= \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \phi} = \sum_{t=1}^n \frac{\partial \ell_t(\boldsymbol{\beta}, \phi)}{\partial \phi} \\ &= \sum_{t=1}^n \left\{ \psi(\phi) - \mu_t \psi(\mu_t \phi) - (1 - \mu_t) \psi((1 - \mu_t) \phi) + \mu_t \log Y_t \right. \\ &\quad \left. + \log(1 - Y_t) - \mu_t \log(1 - Y_t) \right\} \\ &= \sum_{t=1}^n \left\{ \mu_t (Y_t^* - \mu_t^*) + \psi(\phi) - \psi((1 - \mu_t) \phi) + \log(1 - Y_t) \right\} \\ &= \mathbf{1}^\top \mathbf{e}, \end{aligned} \quad (2.7)$$

where $\mathbf{e} = (e_1, \dots, e_n)^\top$, $e_t = \mu_t (Y_t^* - \mu_t^*) + \psi(\phi) - \psi((1 - \mu_t) \phi) + \log(1 - Y_t)$ and $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^n$ (vector of ones of length n).

Altogether, we have the whole score function

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}, \phi) &= \begin{pmatrix} U_{\beta_1}(\boldsymbol{\beta}, \phi) \\ \vdots \\ U_{\beta_k}(\boldsymbol{\beta}, \phi) \\ U_\phi(\boldsymbol{\beta}, \phi) \end{pmatrix} \\ &= \begin{pmatrix} \phi \sum_{t=1}^n (Y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} X_{t1} \\ \vdots \\ \phi \sum_{t=1}^n (Y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} X_{tk} \\ \sum_{t=1}^n \mu_t (Y_t^* - \mu_t^*) + \psi(\phi) - \psi((1 - \mu_t) \phi) + \log(1 - Y_t) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{U}_\beta(\boldsymbol{\beta}, \phi) \\ U_\phi(\boldsymbol{\beta}, \phi) \end{pmatrix} \\ &= \begin{pmatrix} \phi \mathbb{X}^\top \mathbb{T}(\mathbf{Y}^* - \boldsymbol{\mu}^*) \\ \mathbf{1}^\top \mathbf{e} \end{pmatrix}. \end{aligned}$$

2.2.2 Fisher information matrix

We can see that it is not possible to solve $\mathbf{U}(\boldsymbol{\beta}, \phi) = 0$ analytically, and an iterative algorithm is needed to calculate the estimates numerically. Later, we will describe the Fisher scoring algorithm, which requires the Fisher information matrix. The Fisher information matrix will also be needed for parameter inference.

The definition of the Fisher information matrix requires us to calculate the variance of the score function

$$\begin{aligned} I(\boldsymbol{\beta}, \phi) &= \text{var} [\mathbf{U}(\boldsymbol{\beta}, \phi)] = \mathbb{E} [\mathbf{U}(\boldsymbol{\beta}, \phi) \mathbf{U}(\boldsymbol{\beta}, \phi)^\top] \\ &= \mathbb{E} \left[\left(\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\beta}^\top, \phi)^\top} \right) \left(\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\beta}^\top, \phi)^\top} \right)^\top \right], \quad (2.8) \end{aligned}$$

but we will use the alternative formula to derive the Fisher information matrix

$$I(\boldsymbol{\beta}, \phi) = -\mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\beta}^\top, \phi)^\top \partial (\boldsymbol{\beta}^\top, \phi)} \right] = -\mathbb{E} \left[\frac{\partial \mathbf{U}(\boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\beta}^\top, \phi)} \right],$$

which requires the regularity conditions (R5) and (R6) stated in Appendix B to hold. For the beta regression model, they are verified in Section 2.3.1.

Let us denote the blocks of the Fisher information matrix function as

$$I(\boldsymbol{\beta}, \phi) = \begin{pmatrix} I_{\beta\beta}(\boldsymbol{\beta}, \phi) & I_{\beta\phi}(\boldsymbol{\beta}, \phi) \\ I_{\phi\beta}(\boldsymbol{\beta}, \phi) & I_{\phi\phi}(\boldsymbol{\beta}, \phi) \end{pmatrix},$$

where $I_{\beta\beta}(\boldsymbol{\beta}, \phi) = -\mathbb{E} \frac{\partial \mathbf{U}_{\beta}(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}^\top}$, $I_{\beta\phi}(\boldsymbol{\beta}, \phi) = I_{\phi\beta}^\top(\boldsymbol{\beta}, \phi) = -\mathbb{E} \frac{\partial \mathbf{U}_{\beta}(\boldsymbol{\beta}, \phi)}{\partial \phi}$, $I_{\phi\phi}(\boldsymbol{\beta}, \phi) = -\mathbb{E} \frac{\partial \mathbf{U}_{\phi}(\boldsymbol{\beta}, \phi)}{\partial \phi}$.

We have

$$\mathbb{E} \frac{\partial U_{\beta_i}(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \mathbb{E} \frac{\partial U_{\beta_i}(\boldsymbol{\beta}, \phi)}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_j} = \mathbb{E} \left[\frac{\partial U_{\beta_i}(\boldsymbol{\beta}, \phi)}{\partial \mu_t} \right] \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_j},$$

because we know $\frac{\partial \mu_t}{\partial \eta_t} = \frac{1}{g'(\mu_t)}$ and $\frac{\partial \eta_t}{\partial \beta_j} = X_{tj}$. Denote $\mu_t^* = \frac{\partial \mu_t}{\partial \eta_t} = \phi [\psi'(\mu_t \phi) + \psi'((1 - \mu_t) \phi)]$, where $\psi'(\cdot)$ is the trigamma function (see Appendix A). Then

$$\begin{aligned} \mathbb{E} \frac{\partial U_{\beta_i}(\boldsymbol{\beta}, \phi)}{\partial \mu_t} &= \phi \sum_{t=1}^n \mathbb{E} \left[Y_t^* \frac{g''(\mu_t)}{[g'(\mu_t)]^2} X_{ti} - \mu_t^* \frac{1}{g'(\mu_t)} X_{ti} - \mu_t^* \frac{g''(\mu_t)}{[g'(\mu_t)]^2} X_{ti} \right] \\ &= \phi \sum_{t=1}^n \left\{ \mathbb{E} \left[(Y_t^* - \mu_t^*) \frac{g''(\mu_t)}{[g'(\mu_t)]^2} X_{ti} \right] - \mu_t^* \frac{1}{g'(\mu_t)} X_{ti} \right\} \\ &= \phi \sum_{t=1}^n \left\{ \mathbb{E} [(Y_t^* - \mu_t^*)] \frac{g''(\mu_t)}{[g'(\mu_t)]^2} X_{ti} - \mu_t^* \frac{1}{g'(\mu_t)} X_{ti} \right\} \\ &= -\phi \sum_{t=1}^n \mu_t^* \frac{1}{g'(\mu_t)} X_{ti}. \end{aligned}$$

In the last equation, we use $\mathbb{E}(Y_t^*) = \mu_t^*$, which can be verified by direct calculation. Therefore,

$$\mathbb{E} \frac{\partial U_{\beta_i}(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = -\phi \sum_{t=1}^n \mu_t^* \frac{1}{[g'(\mu_t)]^2} X_{ti} X_{tj}.$$

When we denote $w_t = \mu_t^* \frac{1}{[g'(\mu_t)]^2}$ and $\mathbb{W} = \text{diag}(w_1, \dots, w_n)$, we can write

$$I_{\beta\beta}(\boldsymbol{\beta}, \phi) = -\mathbb{E} \frac{\partial \mathbf{U}_{\beta}(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}^\top} = \phi \mathbb{X}^\top \mathbb{W} \mathbb{X}.$$

Now having $c_t = \frac{\partial \mu_t^*}{\partial \phi} = [\psi'(\mu_t \phi) \mu_t - \psi'((1 - \mu_t) \phi)(1 - \mu_t)]$, we get

$$\begin{aligned} \mathbb{E} \frac{\partial U_{\beta_i}(\beta, \phi)}{\partial \phi} &= \mathbb{E} \left[\frac{\partial}{\partial \phi} \left\{ \phi \sum_{t=1}^n (Y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} X_{ti} \right\} \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n (Y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} X_{ti} - \phi \sum_{t=1}^n c_t \frac{1}{g'(\mu_t)} X_{ti} \right] \\ &= \sum_{t=1}^n \mathbb{E} [(Y_t^* - \mu_t^*)] \frac{1}{g'(\mu_t)} X_{ti} - \phi \sum_{t=1}^n c_t \frac{1}{g'(\mu_t)} X_{ti} \\ &= -\phi \sum_{t=1}^n c_t \frac{1}{g'(\mu_t)} X_{ti}, \end{aligned}$$

where in the last equation we used again $\mathbb{E}(Y_t^*) = \mu_t^*$. When we denote $\mathbf{c} = (c_1, \dots, c_n)^\top$, we can write

$$I_{\beta\phi}(\beta, \phi) = I_{\phi\beta}^\top(\beta, \phi) = -\mathbb{E} \frac{\partial \mathbf{U}_\beta(\beta, \phi)}{\partial \phi} = \phi \mathbb{X}^\top \mathbb{T} \mathbf{c}.$$

The remaining part is then calculated as

$$\begin{aligned} \mathbb{E} \frac{\partial U_\phi(\beta, \phi)}{\partial \phi} &= \mathbb{E} \left[\frac{\partial}{\partial \phi} \left\{ \sum_{t=1}^n \mu_t (Y_t^* - \mu_t^*) + \psi(\phi) - \psi((1 - \mu_t) \phi) + \log(1 - Y_t) \right\} \right] \\ &= \sum_{t=1}^n \mathbb{E} [-\mu_t c_t + \psi'(\phi) - (1 - \mu_t) \psi'((1 - \mu_t) \phi)] \\ &= -\sum_{t=1}^n \mu_t^2 \psi'(\mu_t \phi) + (1 - \mu_t)^2 \psi'((1 - \mu_t) \phi) - \psi'(\phi). \end{aligned}$$

Denote $d_t = \mu_t^2 \psi'(\mu_t \phi) + (1 - \mu_t)^2 \psi'((1 - \mu_t) \phi) - \psi'(\phi)$, $\mathbf{d} = (d_1, \dots, d_n)^\top$. Then we have

$$I_{\phi\phi}(\beta, \phi) = -\mathbb{E} \frac{\partial U_\phi(\beta, \phi)}{\partial \phi} = \sum_{t=1}^n d_t = \mathbf{1}^\top \mathbf{d}.$$

Altogether, we have the Fisher information matrix as a block matrix

$$I(\beta, \phi) = \begin{pmatrix} I_{\beta\beta}(\beta, \phi) & I_{\beta\phi}(\beta, \phi) \\ I_{\phi\beta}(\beta, \phi) & I_{\phi\phi}(\beta, \phi) \end{pmatrix} = \begin{pmatrix} \phi \mathbb{X}^\top \mathbb{W} \mathbb{X} & \phi \mathbb{X}^\top \mathbb{T} \mathbf{c} \\ \phi \mathbf{c}^\top \mathbb{T} \mathbb{X} & \mathbf{1}^\top \mathbf{d} \end{pmatrix}. \quad (2.9)$$

Notice that the regression parameters β and the dispersion parameter ϕ are not orthogonal ($I_{\beta\phi}(\beta, \phi) \neq \mathbf{0}$, $I_{\phi\beta}(\beta, \phi)^\top \neq \mathbf{0}$) as in a generalised linear model (MacCullagh and Nelder [1989]). Therefore, β and ϕ have to be estimated together and cannot be calculated separately.

2.2.3 Algorithm for calculation of estimates

As mentioned above, we need to use an iterative algorithm to calculate the estimates. In this section, the Fisher scoring algorithm is described, which requires the inverse of the Fisher information matrix. There are more algorithms; an alternative could be the Newton-Raphson algorithm, described, for example, in Hardin and Hilbe [2007]. To obtain the inverse, we will use the well-known formula (2.10) for the inversion of the block matrix.

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{D}^*\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{D}^* \\ -\mathbf{D}^*\mathbf{C}\mathbf{A}^{-1} & \mathbf{D}^* \end{pmatrix}, \quad (2.10)$$

where $\mathbf{D}^* = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}$. We denote the blocks of the inverse matrix

$$[I(\boldsymbol{\beta}, \phi)]^{-1} = \begin{pmatrix} I^{\beta\beta}(\boldsymbol{\beta}, \phi) & I^{\beta\phi}(\boldsymbol{\beta}, \phi) \\ I^{\phi\beta}(\boldsymbol{\beta}, \phi) & I^{\phi\phi}(\boldsymbol{\beta}, \phi) \end{pmatrix},$$

where $I^{\beta\beta}(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{k \times k}$, $I^{\beta\phi}(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{k \times 1}$, $I^{\phi\beta}(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{1 \times k}$, and $I^{\phi\phi}(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{1 \times 1}$, so we denote it as γ :

$$\begin{aligned} I^{\phi\phi}(\boldsymbol{\beta}, \phi) &= \gamma = (\mathbf{1}^\top \mathbf{d} - \phi \mathbf{c}^\top \mathbf{T} \mathbf{X} (\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \phi \mathbf{X}^\top \mathbf{T} \mathbf{c})^{-1} \\ &= \frac{1}{\mathbf{1}^\top \mathbf{d} - \phi \mathbf{c}^\top \mathbf{T} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{T} \mathbf{c}}. \end{aligned}$$

Now

$$\begin{aligned} I^{\beta\beta}(\boldsymbol{\beta}, \phi) &= (\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} + (\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \phi \mathbf{X}^\top \mathbf{T} \mathbf{c} \gamma \phi \mathbf{c}^\top \mathbf{T} \mathbf{X} (\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \\ &= \frac{1}{\phi} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} + \gamma (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{T} \mathbf{c} \mathbf{c}^\top \mathbf{T} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}, \\ I^{\beta\phi}(\boldsymbol{\beta}, \phi) &= -(\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \phi \mathbf{X}^\top \mathbf{T} \mathbf{c} \gamma \\ &= -\gamma (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{T} \mathbf{c}, \\ I^{\phi\beta}(\boldsymbol{\beta}, \phi) &= -\gamma \phi \mathbf{c}^\top \mathbf{T} \mathbf{X} (\phi \mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \\ &= -\gamma \mathbf{c}^\top \mathbf{T} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} = [I^{\beta\phi}(\boldsymbol{\beta}, \phi)]^\top. \end{aligned}$$

So, we have the whole inverse

$$\begin{aligned} [I(\boldsymbol{\beta}, \phi)]^{-1} &= \begin{pmatrix} I^{\beta\beta}(\boldsymbol{\beta}, \phi) & I^{\beta\phi}(\boldsymbol{\beta}, \phi) \\ I^{\phi\beta}(\boldsymbol{\beta}, \phi) & I^{\phi\phi}(\boldsymbol{\beta}, \phi) \end{pmatrix} = \\ &\begin{pmatrix} \frac{1}{\phi} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} + \gamma (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{T} \mathbf{c} \mathbf{c}^\top \mathbf{T} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} & -\gamma (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{T} \mathbf{c} \\ -\gamma \mathbf{c}^\top \mathbf{T} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} & \gamma \end{pmatrix} \end{aligned}$$

Now we can estimate our regression coefficients using the Fisher scoring algorithm, which is a modification of the Newton algorithm for the optimisation of nonlinear functions.

Algorithm to calculate the estimates of the beta regression coefficients:

1. Set precision $\delta > 0$ for stopping criteria.
2. Define a metric $|\cdot|$ of precision, e.g. L_2 -norm.
3. Set initial values $\boldsymbol{\beta}^{(0)}$, $\phi^{(0)}$ of unknown parameters $\boldsymbol{\beta}$ and ϕ .
4. Set $m = 0$. Until $|(\boldsymbol{\beta}^{(m+1)}, \phi^{(m+1)})^\top - (\boldsymbol{\beta}^{(m)}, \phi^{(m)})^\top| > \delta$ iterate (increment m by 1 at each step):

$$\begin{pmatrix} \boldsymbol{\beta}^{(m+1)} \\ \phi^{(m+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(m)} \\ \phi^{(m)} \end{pmatrix} + [I(\boldsymbol{\beta}^{(m)}, \phi^{(m)})]^{-1} \mathbf{U}(\boldsymbol{\beta}^{(m)}, \phi^{(m)}).$$

5. The maximum likelihood estimates then are

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} = \begin{pmatrix} \beta^{(m+1)} \\ \phi^{(m+1)} \end{pmatrix}.$$

At the beginning of the algorithm, we need some initial values for β and ϕ . For β , it is practical to choose estimated parameters given by linear regression of the link-transformed response on $\mathbb{X}$ as suggested in Ferrari and Cribari-Neto [2004]. We can obtain them by applying ordinary least squares to transformed responses $g(Y_1), \dots, g(Y_n)$, where $g(\cdot)$ is our link function. The initial value of ϕ is suggested in the article as

$$\frac{1}{n} \sum_{t=1}^n \frac{\check{\mu}_t (1 - \check{\mu}_t)}{\check{\sigma}_t^2} - 1, \quad (2.11)$$

where $\check{\mu}_t$ is obtained by applying $g^{-1}(\cdot)$ to the t -th fitted value of the linear regression and $\check{\sigma}_t^2 = \frac{u^\top u}{(n-k)[g'(\check{\mu})]^2}$, where u is the vector of residuals from the applied linear regression. The initial value of ϕ is derived from the variance given by (1.6) and the approximation using the Taylor expansion of $g(Y_t)$ in $\check{\mu}_t$.

2.3 Statistical inference

In this section, we describe the asymptotic properties of our estimates. We briefly discuss consistency, asymptotic distribution, and parameter inference – hypothesis testing and confidence intervals.

2.3.1 Regularity conditions

The asymptotic properties of our maximum likelihood estimates are based on the general maximum likelihood theory, where certain regularity conditions are necessary for the asymptotics to work. In the following, we will verify the conditions stated in the Appendix B for the beta distribution with parameters μ, ϕ .

Let us consider the beta regression model, which can be described as

$$\mathcal{F}_B = \left\{ \text{distributions with density } f_{Y, \mathbf{X}}(y, \mathbf{x}) = f_{Y|\mathbf{X}}(y, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}), \text{ where} \right. \\ \left. f_{Y|\mathbf{X}}(y, \mathbf{x}) \text{ is the density of } B(g^{-1}(\mathbf{x}^\top \beta), \phi), (\beta^\top, \phi)^\top \in R^k \times (0, \infty) \right\},$$

where β, ϕ are beta regression parameters.

The first condition (R1) is satisfied by the definition of the beta regression model, because our model has $k+1$ parameters. The second condition (R2) holds as well; beta distribution density support is $(0, 1)$ regardless of the parameters. $R^k \times (0, \infty)$ is an open set and that verifies condition (R3).

Directly from the shape of (1.5) we can see that the probability density function has continuous derivatives of at least 2 orders (recall the assumption on the link function) with respect to β and ϕ and therefore (R4) is verified.

Conditions (R5) and (R6) hold because the beta distribution belongs to the (general) exponential family of distributions (verified in Section 2.1.2). The validity of these conditions is the consequence of Chapter 1, Theorem 5.8 in Lehmann and Casella [1998], which states that for a distribution of exponential type, the interchange of derivative and integral is valid.

2.3.2 Asymptotic distribution and consistence

It follows from the general maximum likelihood theory (which can be found, for example, in Newey and McFadden [1983]), that the maximum likelihood estimate is consistent under the regularity conditions (R1) – (R6) (see Lehmann and Casella [1998]). Since we verified the regularity conditions for the beta regression model, there exists $n_0 \in \mathbb{N}$ and a sequence of $(\hat{\beta}_{(n)}^\top, \hat{\phi}_{(n)})$ of maximum likelihood estimates defined by (2.2.1) constructed from a sample of $n \geq n_0$ observations such that

$$\begin{pmatrix} \hat{\beta}_{(n)} \\ \hat{\phi}_{(n)} \end{pmatrix} \xrightarrow[n \rightarrow \infty]{P} \begin{pmatrix} \beta^0 \\ \phi^0 \end{pmatrix},$$

where β^0, ϕ^0 are the true values of the parameters and $\xrightarrow{P}$ denotes the convergence in probability.

The other asymptotic result of the maximum likelihood theory is the asymptotic normality of $\hat{\beta}, \hat{\phi}$. Let $(\hat{\beta}_{(n)}^\top, \hat{\phi}_{(n)})$ be the consistent estimate; then

$$\sqrt{n} \left(\begin{pmatrix} \hat{\beta}_{(n)} \\ \hat{\phi}_{(n)} \end{pmatrix} - \begin{pmatrix} \beta^0 \\ \phi^0 \end{pmatrix} \right) \xrightarrow[n \rightarrow \infty]{D} N_{k+1} \left(\mathbf{0}, [I(\beta^0, \phi^0)]^{-1} \right), \quad (2.12)$$

where $N_{k+1}(\cdot, \cdot)$ is $k+1$ -dimensional normal distribution and $\xrightarrow{D}$ denotes convergence in the distribution.

2.3.3 Parameters hypothesis testing

Testing the regression parameters β is important in model building. In this section, the likelihood ratio test, the Rao score test, and the Wald test are described. From a maximum likelihood perspective, they are tests with nuisance parameters, where the tested parameters are some of the components of β .

Let us denote $\beta_A = (\beta_{l_1}, \dots, \beta_{l_m})^\top, l_1, \dots, l_m \in \{1, \dots, k\}, l_i \neq l_j \forall i, j \in \{1, \dots, m, i \neq j\}, l_i < l_j, \forall i < j$ (subvector of vector β). Now, define the hypothesis and the alternative

$$\begin{aligned} \mathcal{H}_0 : \beta_A &= \beta_A^{(0)}, \\ \mathcal{H}_1 : \beta_A &\neq \beta_A^{(0)}, \end{aligned}$$

where $\beta_A^{(0)} = (\beta_{l_1}^{(0)}, \dots, \beta_{l_m}^{(0)})^\top$ is a given vector of constants. The test statistics for the test for $\mathcal{H}_0$ are as follows:

$$\begin{aligned} L &= 2 \log \left[\frac{\mathcal{L}(\hat{\beta}, \hat{\phi})}{\mathcal{L}(\tilde{\beta}, \tilde{\phi})} \right] = 2 \left(\ell(\hat{\beta}, \hat{\phi}) - \ell(\tilde{\beta}, \tilde{\phi}) \right) \\ W &= n \left(\hat{\beta}_A - \beta_A^{(0)} \right)^\top \left[J_{AA}(\hat{\beta}, \hat{\phi}) \right]^{-1} \left(\hat{\beta}_A - \beta_A^{(0)} \right) \\ R &= \frac{1}{n} \mathbf{U}_{\beta_A}(\tilde{\beta}, \tilde{\phi})^\top J_{AA}(\tilde{\beta}, \tilde{\phi}) \mathbf{U}_{\beta_A}(\tilde{\beta}, \tilde{\phi}), \end{aligned}$$

where $\hat{\beta}, \hat{\phi}$ are the model estimates, $\tilde{\beta}, \tilde{\phi}$ are the estimates under the null hypothesis $\mathcal{H}_0$, $J_{AA}(\beta, \phi)$ is the matrix created by taking elements from rows and columns $l_1, \dots, l_m$ of the matrix $[I(\beta, \phi)]^{-1}$, $[J_{AA}(\beta, \phi)]^{-1}$ is the inverse of $J_{AA}(\beta, \phi)$ and

$U_{\beta_A}(\beta, \phi)$ is a vector consisting of elements $l_1, \dots, l_m$ of the model score function $U_{\beta}(\beta, \phi)$.

The test statistic L leads to the likelihood ratio test, W to the Wald test, and R to the Rao score test (in econometrics often called the Lagrange multiplier test). All test statistics converge in distribution to χ_m^2 as $n \rightarrow \infty$ (for a proof, see, for example, Rossi [2018]).

To test the significance of a single regression coefficient ($m = 1$ and $\beta_A^{(0)} = 0$), the square root of the Wald test statistics can be used. That can be seen directly from (2.12) or as a special case of the Wald test described above. The test statistic has a simple form

$$W^* = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)},$$

where $\text{se}(\hat{\beta}_i)$ is the standard error of the estimate, which is a square root of the i -th diagonal element of the matrix $[I(\hat{\beta}, \hat{\phi})]^{-1}$. The test statistic W^* then converges in distribution to the standard normal distribution as $n \rightarrow \infty$.

From the W^* test statistics, we can construct a confidence interval for β_i with a confidence level $1 - \alpha$

$$\left(\hat{\beta}_i - \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\beta}_i), \hat{\beta}_i + \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\beta}_i) \right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Similarly, the confidence interval for ϕ can be derived

$$\left(\hat{\phi} - \Phi^{-1}(1 - \alpha/2) \sqrt{\hat{\gamma}}, \hat{\phi} + \Phi^{-1}(1 - \alpha/2) \sqrt{\hat{\gamma}} \right),$$

where $\hat{\gamma}$ is the estimate of γ with $\hat{\beta}, \hat{\phi}$ substituted.

Another interesting confidence interval is the one for μ_t , which can be derived using a linear combination of $\hat{\beta}$ and $\mathbf{X}_t$ and the fact that $g(\cdot)$ is a monotonous function. If $g(\cdot)$ is increasing, the $(1 - \alpha)$ confidence interval for μ_t is

$$\left(g^{-1} \left(\mathbf{X}_t^\top \hat{\beta} - \Phi^{-1}(1 - \alpha/2) \sqrt{\mathbf{X}_t^\top I^{\beta\beta}(\hat{\beta}, \hat{\phi}) \mathbf{X}_t} \right), \right. \\ \left. g^{-1} \left(\mathbf{X}_t^\top \hat{\beta} + \Phi^{-1}(1 - \alpha/2) \sqrt{\mathbf{X}_t^\top I^{\beta\beta}(\hat{\beta}, \hat{\phi}) \mathbf{X}_t} \right) \right).$$

The confidence region for β_A (subvector of β) can be obtained as a complementary set to the critical region of one of the test statistics L , W , or R described above.

2.4 Model diagnostics

It is important to have tools for the comparison of fitted models. In this section, we describe goodness-of-fit measures and graphs to detect possible defects in the model.

The common measure of explained variation is pseudo R^2 , which can be defined in two ways. In Ferrari and Cribari-Neto [2004], the authors propose to

define R_1^2 as the square of a sample correlation coefficient of the transformed response and the linear predictor

$$R_1^2 = [\widehat{\text{corr}}(Z_t, \hat{\eta}_t)]^2 = \frac{\left[\sum_{t=1}^n (Z_t - \bar{Z})(\hat{\eta}_t - \bar{\eta}) \right]^2}{\sum_{t=1}^n (Z_t - \bar{Z})^2 \sum_{t=1}^n (\hat{\eta}_t - \bar{\eta})^2},$$

where for the link function $g(\cdot)$ we have $Z_t = g(Y_t)$, $\bar{Z} = \frac{1}{n} \sum_{t=1}^n Z_t$, $\hat{\eta}_t = \mathbf{X}_t^\top \hat{\beta}$ and $\bar{\eta} = \frac{1}{n} \sum_{t=1}^n \hat{\eta}_t$.

For the second possible definition, the sample correlation coefficient of the response and the prediction (fitted value) is calculated

$$R_2^2 = [\widehat{\text{corr}}(Y_t, \hat{\mu}_t)]^2 = \frac{\left[\sum_{t=1}^n (Y_t - \bar{Y})(\hat{\mu}_t - \bar{\mu}) \right]^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2 \sum_{t=1}^n (\hat{\mu}_t - \bar{\mu})^2}, \quad (2.13)$$

where $\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$, and $\bar{\mu} = \frac{1}{n} \sum_{t=1}^n \hat{\mu}_t = \sum_{t=1}^n g^{-1}(\hat{\eta}_t)$.

Note that $0 \leq R_1^2 \leq 1$, $0 \leq R_2^2 \leq 1$, because the correlation coefficient has the domain $[-1, 1]$. A perfect model fit (situation, when $\hat{\mu}_t = Y_t, \forall t = 1, \dots, n$) is represented by $R_2^2 = 1$. Because $g(\cdot)$ is monotonous, also $R_2^2 = 1$ for a model with a perfect fit. Generally, a better fit of the model is represented by a higher R_1^2 or R_2^2 .

Another measure of global goodness-of-fit is deviance. Deviance is defined as twice the difference between a saturated model (model with the largest possible number of parameters satisfying the assumption $\text{rank}(\mathbb{X}) = k$) and our investigated model.

$$D(\mathbf{Y}; \hat{\beta}, \hat{\phi}) = 2 \left(\ell(\tilde{\beta}, \tilde{\phi}) - \ell(\hat{\beta}, \hat{\phi}) \right), \quad (2.14)$$

where $\ell(\tilde{\beta}, \tilde{\phi})$ is the log-likelihood in $\tilde{\beta}, \tilde{\phi}$ (estimated parameters in the saturated model). Deviance can be rewritten as a sum of squares of individual deviance residuals

$$D(\mathbf{Y}; \hat{\beta}, \hat{\phi}) = \sum_{t=1}^n (r_t^d)^2, \quad (2.15)$$

where the individual residuals r_t^d are defined as

$$r_t^d = \text{sign}(Y_t - \hat{\mu}_t) \left[2 \left(\ell_t(\tilde{\beta}, \tilde{\phi}) - \ell_t(\hat{\beta}, \hat{\phi}) \right) \right]^{\frac{1}{2}},$$

where $\text{sign}(\cdot)$ is the signum function. Since a lower deviance means a better model fit, we can look at the absolute value of the deviance residuals as indicators of the discrepancy for individual observations. If the deviance residuals are shifted or non-evenly distributed, we can observe possible defects in the model.

Other often used residuals are the standardised (sometimes called Pearson) residuals

$$r_t = \frac{Y_t - \hat{\mu}_t}{\sqrt{\widehat{\text{var}}(Y_t | \mathbf{X}_t)}},$$

where $\widehat{\text{var}}(Y_t | \mathbf{X}_t) = \frac{\hat{\mu}_t(1-\hat{\mu}_t)}{1+\phi}$. Similarly to the deviance residuals, standardised residuals can also be used to identify discrepant observations or other misalignments in the model. Standardised residuals can also be used to identify link

function misspecification. If a wrong link function is chosen, we can see a pattern in the graph of r_t versus $\hat{\eta}_t = \mathbf{X}_t^\top \hat{\boldsymbol{\beta}}$.

For illustration, let $\mathcal{M}_1$ be the true model defined by $Y_t \sim \text{B}(g^{-1}(\eta_t), \phi)$, $t \in \{1, \dots, 200\}$, where $g(\cdot)$ is the logit link function, $\phi = 1000$ (the higher ϕ illustrates the problem clearer), $\eta_t = \beta_1 + X_{t1}\beta_2 + X_{t1}^2\beta_3$, $X_{t1} \sim \text{N}(0, 1)$, $\beta_1 = -2$, $\beta_2 = 1$, and $\beta_3 = 0.5$ are chosen. The model $\mathcal{M}_2$ is then fitted with the omitted regressor X_{1t}^2 and the model $\mathcal{M}_3$ is fitted using the Cauchy link function. In Figures 2.3 and 2.4 we can see the residuals for models $\mathcal{M}_1$ and $\mathcal{M}_2$. We can observe an asymmetry around 0 and more dense observations for negative residuals. An example of the other discrepancy, an inappropriately chosen link function, can be seen in Figure 2.6 (Pearson residuals) and in Figure 2.5 (Deviance residuals), where we compare the logit link model (corresponding to the true data generation) and the model $\mathcal{M}_3$ fitted with the Cauchy link function. From the figures we can clearly see a pattern of the residuals, which indicates an incorrect choice of the link function.

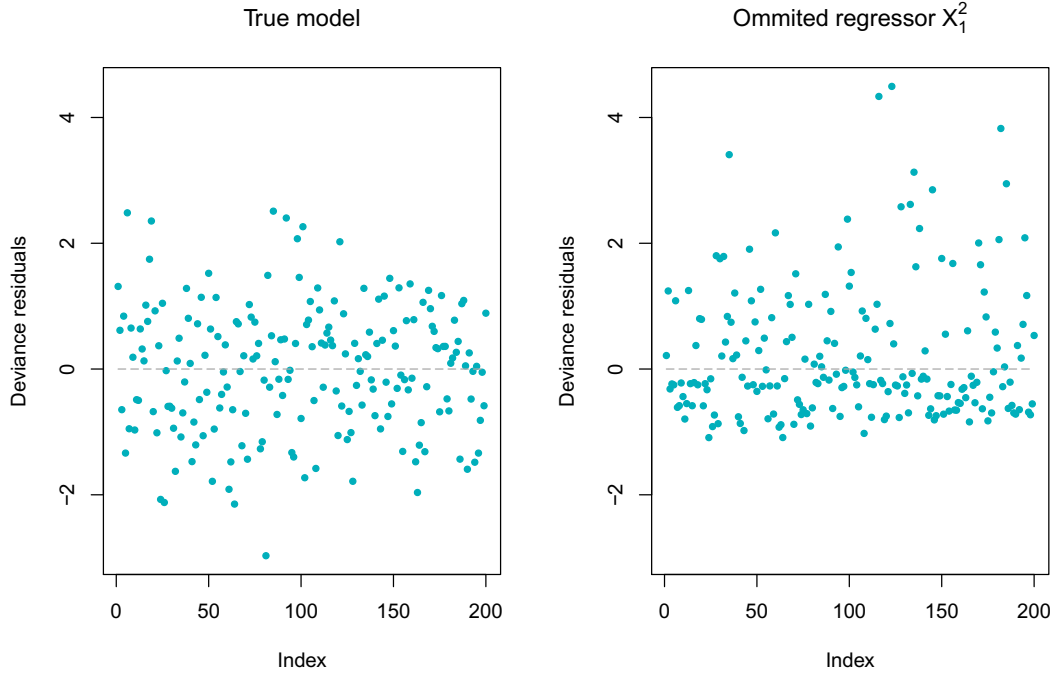


Figure 2.3: Comparison of the deviance residuals for models $\mathcal{M}_1$ (true model) and $\mathcal{M}_2$ (model with omitted regressor X_{t1}^2)

Further research on model discrepancy and influential observations can be found in Ferrari and Cribari-Neto [2004], where generalised influence and Cook's distance are described.

All of the graphs mentioned above can be plotted in R using the package `betareg` with the function `plot`. Instructions and a convenient guide can be found, for example, in Cribari-Neto and Zeileis [2010].

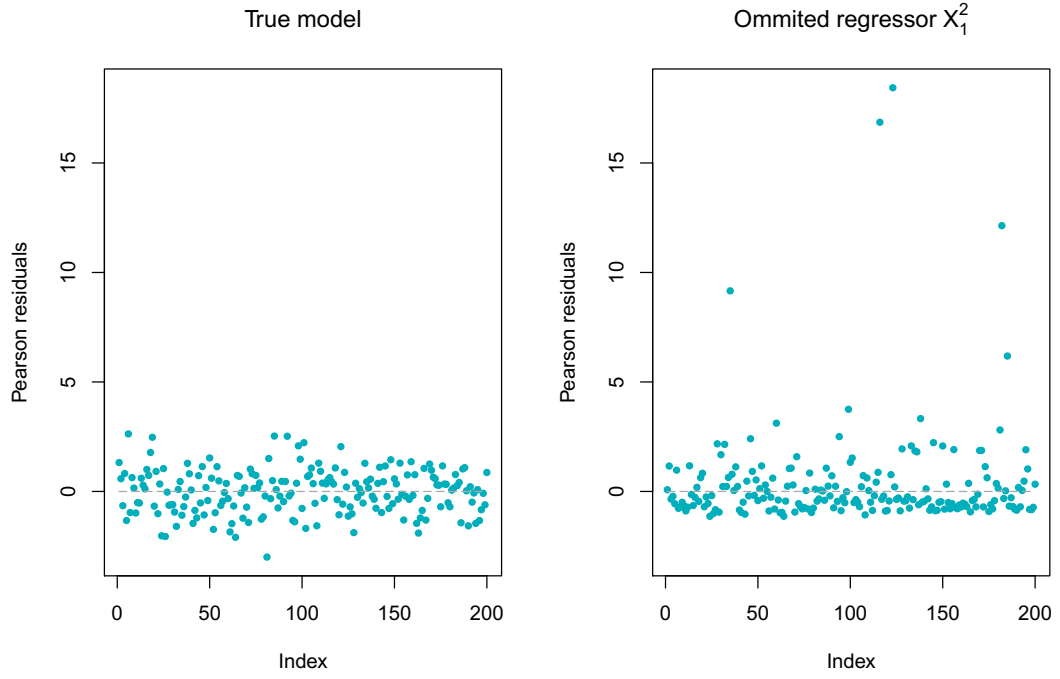


Figure 2.4: Comparison of Pearson residuals for models $\mathcal{M}_1$ (true model) and $\mathcal{M}_2$ (model with omitted regressor X_{t1}^2)

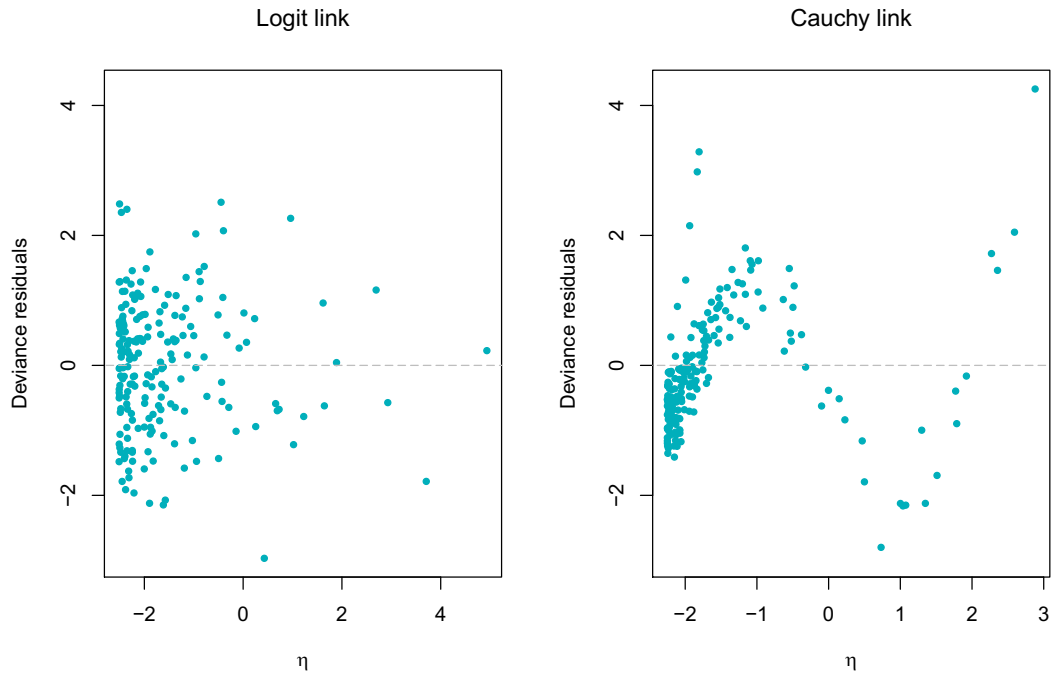


Figure 2.5: Comparison of the deviance residuals against the linear predictor η for models $\mathcal{M}_1$ (true model) and $\mathcal{M}_3$ (model fitted with the Cauchy link function)

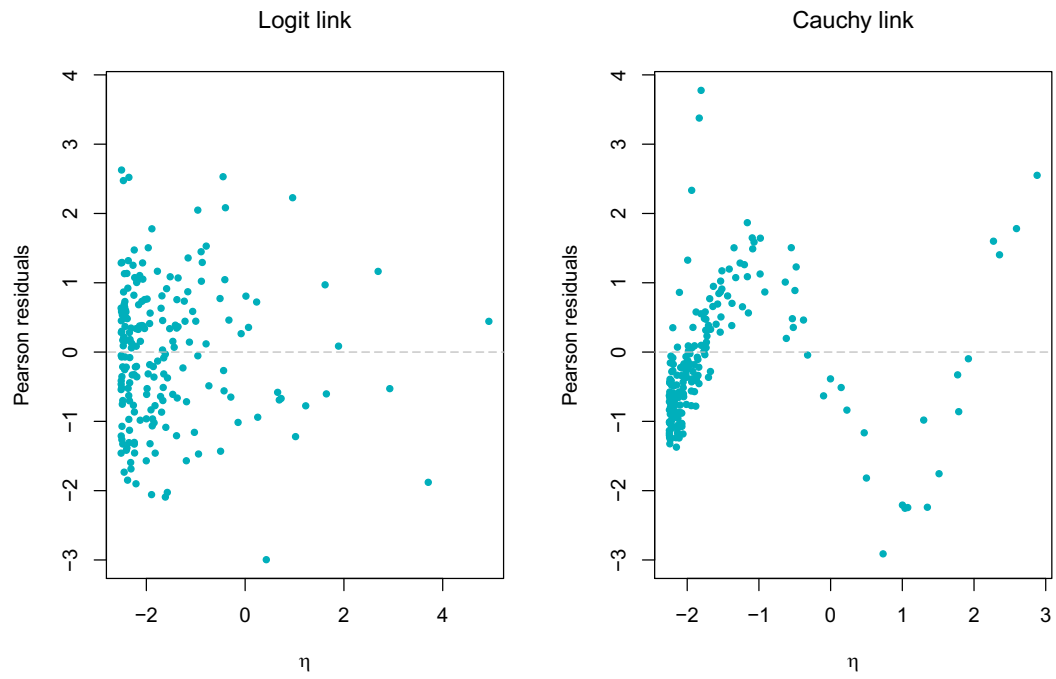


Figure 2.6: Comparison of Pearson residuals against the linear predictor $\boldsymbol{\eta}$ for models $\mathcal{M}_1$ (true model) and $\mathcal{M}_3$ (model fitted with the Cauchy link function)

3. Extension to 0 or 1

In many real-life applications, distribution domains are not the open interval $(0, 1)$, but the random variable can also take a value in one of the boundaries, 0 or 1. For example, if we investigate the mortality rate of car accidents within districts in a state, it is very likely that for a certain district and age category of the driver, we will observe only car accidents without casualties. On the other hand, if the variable examined is an employment rate, we can find districts where the employment rate is 100%.

The beta regression model defined in Definition 3 requires a conditional distribution $Y_t|\mathbf{X}_t \sim B(\mu_t, \phi), t = 1, \dots, n$, which implies that the range of values of $Y_t|\mathbf{X}_t$ is $(0, 1)$, but as mentioned above, in applications we can sometimes observe $Y_t = 0$ or $Y_t = 1$. Not only because of the assumption of the beta distribution, but also because $Y_t = 0$ or $Y_t = 1$ cannot be put into the log-likelihood (2.5), the beta regression is not suitable for modelling this type of data.

In the next section, we summarise three simple approaches to such data often used by practitioners. All of the approaches focus on adjusting the data to fit the beta regression model. An alternative to those simple approaches is the 0-or-1 inflated beta regression model (in this work it is called c -inflated beta regression model) described in Ospina and Ferrari [2012]. In Section 3.2 we properly define the model, derive the maximum likelihood estimates of the coefficients in more detail than in the article, briefly describe statistical inference for the parameters and propose global goodness-of-fit measures and diagnostic plots. The last section is dedicated to the description of predictions for the simple approaches and for the c -inflated beta regression model.

In this work, we will consider only the case where only one of the boundary values is possible for the response. For cases where both 0 and 1 are possible, the methods below can simply be extended and the theory of the 0-and-1 inflated beta regression model can be derived similarly (see Abdel-Karim [2017], for example).

A comparison of all four suggested treatments can be found in a simulation study in Section 4.2.

3.1 Simple approaches

Let us have data $(Y_t, \mathbf{X}_t^\top)^\top, t = 1, \dots, n$, which are assumed to be independent identically distributed random vectors, where $Y_t \in (0, 1) \cup \{c\}, t = 1, \dots, n$, where $c = 0$ or $c = 1$ (the same for all Y_t). We would like to find ways to use the beta regression model with regressors $\mathbf{X}_t$ and response Y_t .

3.1.1 Omitting observations with $Y_t = c$

The first and often used solution by practitioners is to omit observations such that $Y_t = c$. After that, the data (with an appropriate choice of a link function) can be investigated using the beta regression model defined in Definition 3.

The advantage of this approach is mainly its simplicity. The method has several disadvantages; mainly, it reduces the number of observations and, therefore, increases the variability of the parameters. The estimates can be biased

because we do not omit observations at random but only the boundary values. In addition, the number of observations is now a random variable; therefore, more assumptions are needed to ensure that maximum likelihood theory works as expected. Clearly, the disadvantages are more serious with a larger portion of the observations being 0 or 1, therefore, this approach should be applied only to samples with a very low proportion of the boundary values.

3.1.2 Adding/subtracting ε

The second approach is to add (if we have observations where $Y_t = 0$) or subtract (if $Y_t = 1$) a small positive constant ε to such observations to keep the response values in the interval $(0, 1)$. That is

$$Y_t^* = \begin{cases} Y_t & \text{if } Y_t \in (0, 1), \\ Y_t - \text{sign}(Y_t - \frac{1}{2})\varepsilon & \text{if } Y_t \in \{0, 1\}, \end{cases}$$

where $\text{sign}(\cdot)$ is the signum function. Then the beta regression model for the modified data $(Y_t^*, \mathbf{X}_t^\top)^\top$ can be used.

The advantage of this method is that it does not omit observations, and therefore there is lower variability within the estimates. However, with a larger portion of the observations having the boundary value, the estimates (and therefore the fitted values) are more biased because the shift influences the mean.

3.1.3 Multiplying each Y_t

The other possibility is to “shrink” the values of Y_t to be in $(0, 1)$. This can be done, for example, using the following formula

$$Y_t^* = \frac{Y_t - \frac{1}{2}}{1 + \varepsilon} + \frac{1}{2}, \quad (3.1)$$

where $\varepsilon > 0$ is a small constant. In the formula, we subtract $\frac{1}{2}$ to have the domain $(-\frac{1}{2}, \frac{1}{2})$, then shrink it by dividing the value by $1 + \varepsilon$ and add $\frac{1}{2}$. Then we can use the beta regression model for the transformed data $(Y_t^*, \mathbf{X}_t^\top)^\top$.

This approach is different from the two above approaches because we transform each $Y_t, t \in \{1, \dots, n\}$. We can see that $E Y_t^* = \frac{E Y_t - \frac{1}{2}}{1 + \varepsilon} + \frac{1}{2}$ and $\text{var}(Y_t^*) = \frac{1}{(1 + \varepsilon)^2} \text{var}(Y_t)$. The shrinkage influences the dispersion parameter (and therefore the variance of Y_t) rather than the conditional mean. Especially, for $\mu_t = \frac{1}{2}$, we have $E Y_t^* = E Y_t$.

The method is suitable for situations where the mean is not close to the boundary values and we do not insist on a precise estimate of the parameter ϕ .

3.2 Inflated beta regression model

An approach which we will investigate in further detail is based on mixture distributions. We will assume that the random variable Y_t conditionally follows a mixture distribution, where one of the distributions is a beta distribution and the other is the Dirac measure in c ($c = 0$ or $c = 1$). The major advantage of

this approach is that we are modelling the original data and no transformation is needed. On the other hand, a more complex theory is needed, and the estimates have to be constructed in two steps.

In this section, the c -inflated beta regression model is defined and its likelihood, score function, and algorithm for estimation of the coefficients are described. Diagnostic methods for the model are also briefly discussed at the end of this section. Later, in Section 4.2, the proposed approaches are compared for the experimental data. The definition and this section are inspired by Ospina and Ferrari [2012], where the authors propose a model in which all parameters are dependent on regressors. We will focus on a model where ϕ has the same value regarding the regressors, as in Definition 3.

3.2.1 Model definition

First, let us define the c -inflated beta distribution.

Definition 4. *Let $c = 0$ or $c = 1$. The random variable Y^c follows a c -inflated beta distribution, if its probability density function with respect to the sum of the Lebesgue measure and the Dirac measure at c can be written as*

$$f_{Y^c}(y) = \begin{cases} \alpha, & \text{if } y = c, \\ (1 - \alpha)f_Y(y), & \text{if } y \in (0, 1), \end{cases} \quad (3.2)$$

where $\alpha \in (0, 1)$ and $f_Y(y)$ is the probability density function of the beta distribution $B(\mu, \phi)$ given by (1.1). We will denote the random variable Y^c following the c -inflated beta distribution with parameters α, μ, ϕ as $Y^c \sim B^c(\alpha, \mu, \phi)$.

It can easily be seen that $f_{Y^c}(y)$ is a probability density function, because it consists of two probability density functions with weights normalising its integral to 1.

We can rewrite (3.2) in the following way

$$f_{Y^c}(y) = \alpha^{\mathbb{1}_c(y)} (1 - \alpha)^{(1 - \mathbb{1}_c(y))} [f_Y(y)]^{(1 - \mathbb{1}_c(y))}, \quad (3.3)$$

where $\mathbb{1}_c(y)$ is an indicator if $y = c$. This formula will be later used in the derivation of the maximum likelihood estimates of the parameters. Also, we can directly see the mean of Y^c

$$\mathbb{E} Y^c = \alpha c + (1 - \alpha)\mu, \quad (3.4)$$

and calculate the variance

$$\text{var } Y^c = (1 - \alpha) \frac{\mu(1 - \mu)}{1 + \phi} + \alpha(1 - \alpha)(c - \mu)^2.$$

Note that $\mathbb{E} [Y^c | Y^c \in (0, 1)] = \mu$ and $\text{var}[Y^c | Y^c \in (0, 1)] = \frac{\mu(1 - \mu)}{1 + \phi}$. More properties can be found in Ospina and Ferrari [2008].

Now, let us proceed to the definition of the c -inflated beta regression model.

Definition 5. *The data $(Y_t^c, \mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top, t = 1, \dots, n$ satisfy the c -inflated beta regression model if*

1. $(Y_t^c, \mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top, t = 1, \dots, n$ are independent identically distributed random vectors and $P(\text{rank}(\mathbb{X}) = k) = 1, P(\text{rank}(\mathbb{V}) = m) = 1$, where $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^\top$ and $\mathbb{V} = (\mathbf{V}_1, \dots, \mathbf{V}_n)^\top$.
2. Conditional distribution $Y_t^c | \mathbf{X}_t, \mathbf{V}_t \sim B^c(\alpha_t, \mu_t, \phi), t = 1, \dots, n$, where $\alpha_t, \mu_t, t \in \{1, 2, \dots, n\}$ and ϕ are unknown parameters.
3. For all $t = 1, \dots, n$ μ_t depends on $\mathbf{X}_t$ through linear predictor $\eta_t = \mathbf{X}_t^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^\top$ are unknown regression parameters.
4. There exists strictly monotonous, twice differentiable function $g(\cdot)$ mapping $(0, 1)$ to $\mathbb{R}$ which links conditional means μ_t and linear predictors η_t

$$g(\mu_t) = \eta_t = \sum_{i=1}^k X_{ti} \beta_i = \mathbf{X}_t^\top \boldsymbol{\beta}.$$

5. For all $t = 1, \dots, n$ α_t depends on $\mathbf{V}_t$ through linear predictor $\zeta_t = \mathbf{V}_t^\top \boldsymbol{\rho}$, where $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_m)^\top$ are unknown regression parameters.
6. There exists strictly monotonous, twice differentiable function $h(\cdot)$ mapping $(0, 1)$ to $\mathbb{R}$ which links parameter α_t and linear predictors ζ_t

$$h(\alpha_t) = \zeta_t = \sum_{i=1}^m V_{ti} \rho_i = \mathbf{V}_t^\top \boldsymbol{\rho}.$$

where $c = 0$ or $c = 1$.

In the definition of the c -inflated beta regression model, there are two sets of covariates: $\mathbf{X}_t$ and $\mathbf{V}_t$. These covariates can overlap, which means that $X_{ti} = V_{tj}, \forall t \in \{1, \dots, n\}$ for some $i \in \{1, \dots, k\}, j \in \{1, \dots, m\}$. In practical applications, they are often the same or overlap in most elements. However, the definition allows them also to be separated.

3.2.2 Coefficient estimates

To estimate the coefficients of the c -inflated beta regression model, we will follow the maximum likelihood theory, similarly as in Chapter 2. We want to find a maximum likelihood estimate of $\boldsymbol{\theta} = (\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \phi)^\top, \boldsymbol{\theta} = \mathbb{R}^m \times \mathbb{R}^k \times (0, \infty)$ as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \mathcal{L}^*(\boldsymbol{\theta}),$$

where $\mathcal{L}^*(\boldsymbol{\theta})$ is the likelihood function of our data.

Let $(Y_t^c, \mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top, t = 1, \dots, n$ be data that satisfy the definition of the c -inflated beta regression model (Definition 5) with true parameters $\boldsymbol{\rho}^0, \boldsymbol{\beta}^0, \phi^0$. The full likelihood of the data is the following

$$\begin{aligned} \mathcal{L}^*(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \prod_{t=1}^n f_{Y^c, \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t) = \prod_{t=1}^n f_{Y^c | \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t) f_{\mathbf{X}, \mathbf{V}}(\mathbf{X}_t, \mathbf{V}_t) \\ &= \left(\prod_{t=1}^n f_{Y^c | \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t) \right) \left(\prod_{t=1}^n f_{\mathbf{X}, \mathbf{V}}(\mathbf{X}_t, \mathbf{V}_t) \right), \end{aligned}$$

where $f_{Y^c, \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t)$ is the joint distribution of $(Y_t^c, \mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top$, the density $f_{Y^c | \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t)$ is the conditional density of $Y_t^c | \mathbf{X}_t, \mathbf{V}_t \sim \text{B}^c(\alpha_t, \mu_t, \phi)$ and $f_{\mathbf{X}, \mathbf{V}}(\mathbf{X}_t, \mathbf{V}_t)$ is the joint density of $(\mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top$.

Since we want to maximise $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$ with respect to $\boldsymbol{\rho}, \boldsymbol{\beta}, \phi$, we only need to maximise

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \prod_{t=1}^n f_{Y^c | \mathbf{X}, \mathbf{V}}(Y_t^c, \mathbf{X}_t, \mathbf{V}_t).$$

From (3.3) we can write

$$f_{Y^c | \mathbf{X}, \mathbf{V}}(y, \mathbf{x}_t, \mathbf{v}_t) = \alpha_t^{\mathbb{1}_c(y)} (1 - \alpha_t)^{(1 - \mathbb{1}_c(y))} [f_Y(y)]^{(1 - \mathbb{1}_c(y))}, \quad (3.5)$$

where $f_Y(y)$ is the density of the beta distribution $\text{B}(g^{-1}(\mathbf{x}_t^\top \boldsymbol{\beta}), \phi)$ and $\mathbb{1}_c(y)$ is an indicator if $y = c$. Denote $M_t = \mathbb{1}_c(Y_t^c)$ the random variable defined as an indicator if $Y_t^c = c$ and $L = \{t \in \{1, \dots, n\} : M_t = 0\} = \{t \in \{1, \dots, n\} : Y_t^c \in (0, 1)\}$ the set of indexes corresponding to $Y_t^c \neq c$. Therefore, the likelihood

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \prod_{t=1}^n \alpha_t^{M_t} (1 - \alpha_t)^{(1 - M_t)} [f_Y(Y_t)]^{(1 - M_t)} = \prod_{t=1}^n \alpha_t^{M_t} (1 - \alpha_t)^{(1 - M_t)} \prod_{t \in L} f_Y(Y_t),$$

can be factorised as $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \mathcal{L}(\boldsymbol{\rho}) \mathcal{L}(\boldsymbol{\beta}, \phi)$, where

$$\begin{aligned} \mathcal{L}(\boldsymbol{\rho}) &= \prod_{t=1}^n \alpha_t^{M_t} (1 - \alpha_t)^{(1 - M_t)}, \\ \mathcal{L}(\boldsymbol{\beta}, \phi) &= \prod_{t \in L} f_Y(y). \end{aligned}$$

Because the first term depends only on $\boldsymbol{\rho}$ and the second term depends only on $\boldsymbol{\beta}$ and ϕ , the parameters are separable (Pace and Salvani [1997]).

Now, we proceed to the log-likelihood with $\ell(\boldsymbol{\rho}) = \log(\mathcal{L}(\boldsymbol{\rho}))$ and $\ell(\boldsymbol{\beta}, \phi) = \log(\mathcal{L}(\boldsymbol{\beta}, \phi))$. We can write

$$\ell(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \ell(\boldsymbol{\rho}) + \ell(\boldsymbol{\beta}, \phi) = \sum_{t=1}^n \ell_t(\boldsymbol{\rho}) + \sum_{t \in L} \ell_t(\boldsymbol{\beta}, \phi),$$

where

$$\ell_t(\boldsymbol{\rho}) = M_t \log(\alpha_t) + (1 - M_t) \log(1 - \alpha_t), \quad t = 1, \dots, n, \quad (3.6)$$

$$\begin{aligned} \ell_t(\boldsymbol{\beta}, \phi) &= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1 - \mu_t) \phi) + \mu_t \phi \log Y_t^c - \log Y_t^c \\ &\quad + \phi \log(1 - Y_t^c) - \mu_t \phi \log(1 - Y_t^c) - \log(1 - Y_t^c), \quad t \in L. \end{aligned} \quad (3.7)$$

Notice that (3.7) is the same as (2.4) with the only difference that now we take only those $t \in \{1, \dots, n\}$ for which $Y_t^c \in (0, 1)$.

As before, we would like to find the maximum likelihood estimate as a solution of the equations

$$\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \frac{\partial \ell(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial (\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \phi)^\top} = \mathbf{0}. \quad (3.8)$$

Denote the elements of the score function corresponding to $\boldsymbol{\rho}, \boldsymbol{\beta}, \phi$ as

$$\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = (\mathbf{U}_\rho(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)^\top, \mathbf{U}_\beta(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)^\top, U_\phi(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi))^\top,$$

where

$$\begin{aligned} \mathbf{U}_\rho(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= (U_{\rho_1}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi), \dots, U_{\rho_m}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi))^\top, \\ \mathbf{U}_\beta(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= (U_{\beta_1}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi), \dots, U_{\beta_k}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi))^\top. \end{aligned}$$

Denote $\mathbb{M} = \text{diag}(1 - M_1, \dots, 1 - M_n)$, then we can use separability to rewrite $\mathbf{U}_\beta(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$ and $U_\phi(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$ using the results (2.6) and (2.7)

$$\begin{aligned} \mathbf{U}_\beta(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \phi \mathbb{X}^\top \mathbb{T} \mathbb{M} (\mathbf{Y}^{*c} - \boldsymbol{\mu}^*) \\ U_\phi(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \mathbf{1}^\top \mathbb{M} \mathbf{e} \end{aligned}$$

where we set $\mathbf{Y}^{*c} = (Y_1^{*c}, \dots, Y_n^{*c})^\top$, $Y_t^{*c} = Y_t^*$ if $Y_t \in (0, 1)$, $Y_t^{*c} = 0$ otherwise, and $\boldsymbol{\mu}^*$, $\mathbb{T}$ are defined analogously to those in Section 2.2.

For $\mathbf{U}_\rho(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$ we write

$$\mathbf{U}_{\rho_j}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \frac{\partial \ell(\boldsymbol{\rho})}{\partial \rho_j} = \sum_{t=1}^n \frac{\partial \ell_t(\boldsymbol{\rho})}{\partial \alpha_t} \frac{\partial \alpha_t}{\partial \zeta_t} \frac{\partial \zeta_t}{\partial \rho_j}.$$

Knowing $\frac{\partial \alpha_t}{\partial \zeta_t} = \frac{1}{h'(\alpha_t)}$ and $\frac{\partial \zeta_t}{\partial \rho_j} = V_{tj}$, we can calculate the remaining part

$$\frac{\partial \ell_t(\boldsymbol{\rho})}{\partial \alpha_t} = \frac{M_t}{\alpha_t} - \frac{1 - M_t}{1 - \alpha_t}.$$

If we take $\mathbf{m} = (m_1, \dots, m_n)^\top$, $m_t = \frac{M_t}{\alpha_t} - \frac{1 - M_t}{1 - \alpha_t}$, $\mathbb{S} = \text{diag}(\frac{1}{h'(\alpha_1)}, \dots, \frac{1}{h'(\alpha_n)})$, we have the following

$$\mathbf{U}_\rho(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \mathbb{V}^\top \mathbb{S} \mathbf{m}.$$

Altogether, we have the whole score function for the c -inflated beta regression model is

$$\mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \begin{pmatrix} \mathbf{U}_\rho(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \\ \mathbf{U}_\beta(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \\ U_\phi(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \end{pmatrix} = \begin{pmatrix} \mathbb{V}^\top \mathbb{S} \mathbf{m} \\ \phi \mathbb{X}^\top \mathbb{T} \mathbb{M} (\mathbf{Y}^{*c} - \boldsymbol{\mu}^*) \\ \mathbf{1}^\top \mathbb{M} \mathbf{e} \end{pmatrix}.$$

To derive the Fisher information matrix, we can use the alternative formula as in Section 2.2; from (3.5) we can see that the density is the product of a probability density function of the alternative distribution and a probability density function of the beta distribution modified with the exponent $1 - \mathbb{1}_c(y)$ (so it is equal to the density of the beta distribution or 1), which are both members of the exponential family of distributions. Therefore, their product also has an exponential-type distribution and thanks to Theorem 5.8 in Chapter 1 in Lehmann and Casella [1998], the regularity conditions (R5) and (R6) from Appendix B are met and the alternative formula can be used.

We want to calculate

$$I(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = -\mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial(\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \phi)^\top \partial(\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \phi)} \right] = -\mathbb{E} \left[\frac{\partial \mathbf{U}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial(\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \phi)} \right],$$

and we denote similarly as in Section 2.2 blocks of the matrix corresponding to the parameters as

$$I(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \begin{pmatrix} I_{\rho\rho}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\rho\beta}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\rho\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \\ I_{\beta\rho}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\beta\beta}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\beta\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \\ I_{\phi\rho}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\phi\beta}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\phi\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \end{pmatrix}.$$

Thanks to the separability of the parameters $\boldsymbol{\rho}$ and $(\boldsymbol{\beta}^\top, \phi)^\top$, we know that

$$\begin{aligned} I_{\boldsymbol{\rho}\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= I_{\boldsymbol{\beta}\boldsymbol{\rho}}^\top(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \mathbf{0}, \\ I_{\boldsymbol{\rho}\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= I_{\phi\boldsymbol{\rho}}^\top(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \mathbf{0}, \end{aligned}$$

where $\mathbf{0} \in \mathbb{R}^{m \times k}$, $\mathbf{0} \in \mathbb{R}^m$ are the null matrix and the null vector. Also, the blocks corresponding to $\boldsymbol{\beta}, \phi$ are the same as in the beta regression model when we use the modification by $\mathbb{M}$

$$\begin{aligned} I_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \phi \mathbb{X}^\top \mathbb{W} \mathbb{M} \mathbb{X}, \\ I_{\boldsymbol{\beta}\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= I_{\phi\boldsymbol{\beta}}^\top(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \phi \mathbb{X}^\top \mathbb{T} \mathbb{M} \mathbf{c}, \\ I_{\phi\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \mathbf{1}^\top \mathbb{M} \mathbf{d}. \end{aligned}$$

Then only the upper left block remains to be calculated. We have

$$\mathbb{E} \frac{\partial U_{\rho_i}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \rho_j} = \mathbb{E} \frac{\partial U_{\rho_i}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \alpha_t} \frac{\partial \alpha_t}{\partial \zeta_t} \frac{\partial \zeta_t}{\partial \rho_j} = \mathbb{E} \left[\frac{\partial U_{\rho_i}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \alpha_t} \right] \frac{\partial \alpha_t}{\partial \zeta_t} \frac{\partial \zeta_t}{\partial \rho_j},$$

and we know $\frac{\partial \alpha_t}{\partial \zeta_t} = \frac{1}{h'(\alpha_t)}$ and $\frac{\partial \zeta_t}{\partial \rho_j} = V_{tj}$. Proceeding by denoting $m'_t = \frac{\partial m_t}{\partial \alpha_t} = -\left(\frac{M_t}{\alpha_t^2} + \frac{1-M_t}{(1-\alpha_t)^2}\right)$, we have

$$\begin{aligned} \mathbb{E} \frac{\partial U_{\rho_i}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \alpha_t} &= \sum_{t=1}^n \mathbb{E} \left[m'_t \frac{1}{h'(\alpha_t)} + m_t \frac{h''(\alpha_t)}{[h'(\alpha_t)]^2} \right] V_{ti} \\ &= \sum_{t=1}^n \left[\mathbb{E}[m'_t] \frac{1}{h'(\alpha_t)} + \mathbb{E}[m_t] \frac{h''(\alpha_t)}{[h'(\alpha_t)]^2} \right] V_{ti} \\ &= \sum_{t=1}^n \left[-\left(\frac{1}{\alpha_t} + \frac{1}{1-\alpha_t} \right) \frac{1}{h'(\alpha_t)} \right] V_{ti} \\ &= -\sum_{t=1}^n \frac{1}{\alpha_t(1-\alpha_t)} \frac{1}{h'(\alpha_t)} V_{ti}, \end{aligned}$$

where we used $\mathbb{E} M_t = \mathbb{E} \mathbb{1}_c(Y_t^c) = P(Y_t^c = c) = \alpha_t$. Therefore

$$\mathbb{E} \frac{\partial U_{\rho_i}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \rho_j} = -\sum_{t=1}^n \frac{1}{\alpha_t(1-\alpha_t)} \frac{1}{[h'(\mu_t)]^2} V_{ti} V_{tj}.$$

When we denote $a_t = \frac{1}{\alpha_t(1-\alpha_t)}$ and $\mathbb{A} = \text{diag}(a_1, \dots, a_n)$, we can write

$$I_{\boldsymbol{\rho}\boldsymbol{\rho}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) = \mathbb{E} \frac{\partial U_{\boldsymbol{\rho}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\rho}} = \mathbb{V}^\top \mathbb{A} \mathbb{S}^2 \mathbb{V}.$$

Altogether, the Fisher information matrix of the c -inflated beta regression model is

$$\begin{aligned} I(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) &= \begin{pmatrix} I_{\boldsymbol{\rho}\boldsymbol{\rho}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^\top & I_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\boldsymbol{\beta}\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \\ \mathbf{0}^\top & I_{\phi\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) & I_{\phi\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{V}^\top \mathbb{A} \mathbb{S}^2 \mathbb{V} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^\top & \phi \mathbb{X}^\top \mathbb{W} \mathbb{M} \mathbb{X} & \phi \mathbb{X}^\top \mathbb{T} \mathbb{M} \mathbf{c} \\ \mathbf{0}^\top & \phi \mathbf{c}^\top \mathbb{M} \mathbb{T} \mathbb{X} & \mathbf{1}^\top \mathbb{M} \mathbf{d} \end{pmatrix}. \end{aligned}$$

Since we cannot find a solution of (3.8) analytically, we will use the Fisher scoring algorithm. In this case, we can use separability of the parameters to fit the model in 2 steps

1. Use Fisher scoring algorithm for estimating the coefficients $\boldsymbol{\rho} = (\rho_1, \dots, \rho_m)$.
2. Take as data only observations of $(Y_t^c, \mathbf{X}_t^\top, \mathbf{V}_t^\top)^\top$ with $Y_t^c \neq c \iff Y_t^c \in (0, 1)$ and run the algorithm for $\boldsymbol{\beta}, \phi$ as in Section 2.2.3.

Step 2 is equivalent to estimating the coefficients using $I_{\beta\phi}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$, the lower right block of the matrix $I(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$ corresponding to the coefficients $\boldsymbol{\beta}, \phi$, because the matrix $\mathbb{M}$ ensures that we calculate the matrix using only the observations for which $Y_t^c \neq c$.

Let us note here that the first step (calculating the estimates of $\boldsymbol{\rho}$) is equivalent to fitting a generalised linear model for binary data, where the response is $M_t = \mathbb{1}_c(Y_t^c) \sim \text{Alt}(P(Y_t^c = c)) = \text{Alt}(\alpha_t)$ and therefore also the iterative weighted least squares algorithm can be used (see, for example, Bjorck [1996]).

3.2.3 Statistical inference

Statistical tests for parameters $\boldsymbol{\beta}, \phi$ and μ_t are the same as in 2.3, when we take data without observations for which $Y_t^c = c$. For inference on $\boldsymbol{\rho}$, we can follow the same approach as in 2.3.3 and derive the likelihood ratio test, the Wald test, and the Rao test using the inverse of $I_{\rho\rho}(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$. To test a joint hypothesis about $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$, we can use the maximum likelihood theory and derive the three tests using the inverse of the entire Fisher information matrix $I(\boldsymbol{\rho}, \boldsymbol{\beta}, \phi)$.

The $(1 - \alpha)$ confidence interval for ρ_i can be derived from the Wald test as well

$$\left(\hat{\rho}_i - \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\rho}_i), \hat{\rho}_i + \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\rho}_i) \right),$$

where $\text{se}(\hat{\rho}_i)$ is the standard error of the estimate, which is the square root of the i -th diagonal element of the matrix $\left[I_{\rho\rho}(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}, \hat{\phi}) \right]^{-1}$. What is also interesting is the confidence interval for $\mu_t^c = E[Y_t^c | \mathbf{X}_t, \mathbf{V}_t] = \alpha_t c + (1 - \alpha_t) \mu_t$ proposed in Ospina and Ferrari [2012]

$$\left(\hat{\mu}_t^c - \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\mu}_t^c), \hat{\mu}_t^c + \Phi^{-1}(1 - \alpha/2) \text{se}(\hat{\mu}_t^c) \right), \quad (3.9)$$

where $\hat{\mu}_t^c = \hat{\alpha}_t c + (1 - \hat{\alpha}_t) \hat{\mu}_t$ with $\hat{\alpha}_t = h^{-1}(\mathbf{V}^\top \hat{\boldsymbol{\rho}})$, $\hat{\mu}_t = g^{-1}(\mathbf{X}^\top \hat{\boldsymbol{\beta}})$ and

$$\text{se}(\hat{\mu}_t^c) = \sqrt{\left[\frac{c - \hat{\mu}_t}{h'(\hat{\alpha}_t)} \right]^2 \mathbf{V}_t^\top \left[I_{\rho\rho}(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}, \hat{\phi}) \right]^{-1} \mathbf{V}_t + \left[\frac{1 - \hat{\alpha}_t}{g'(\hat{\mu}_t)} \right]^2 \mathbf{X}_t^\top \left[I_{\beta\beta}(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}, \hat{\phi}) \right]^{-1} \mathbf{X}_t}.$$

3.2.4 Model diagnostics

Firstly, we are interested in goodness-of-fit measures. In this case, we have pseudo R^2 for both continuous and discrete parts of the model, and also for the whole model. For the continuous part of the model, we define the pseudo R^2 as

$$R_{2,c}^2 = \frac{\left[\sum_{t \in L} (Y_t^c - \bar{Y}^c) (\hat{\mu}_t - \bar{\mu}) \right]^2}{\sum_{t \in L} (Y_t^c - \bar{Y}^c)^2 \sum_{t \in L} (\hat{\mu}_t - \bar{\mu})^2},$$

where $\bar{Y}^c = \frac{1}{n} \sum_{t \in L} Y_t^c$ and $\bar{\mu} = \frac{1}{n} \sum_{t \in L} \hat{\mu}_t$. Notice that it corresponds to (2.13) calculated for $t \in L$ only. For the discrete part of the model, we can similarly define

$$R_{2,d}^2 = [\widehat{\text{corr}}(M_t, \hat{\alpha}_t)]^2 = \frac{\left[\sum_{t=1}^n (M_t - \bar{M}) (\hat{\alpha}_t - \bar{\alpha}) \right]^2}{\sum_{t=1}^n (M_t - \bar{M})^2 \sum_{t=1}^n (\hat{\alpha}_t - \bar{\alpha})^2},$$

where $\bar{M} = \frac{1}{n} \sum_{t=1}^n M_t$ and $\bar{\alpha} = \frac{1}{n} \sum_{t=1}^n \hat{\alpha}_t$.

We define the global goodness-of-fit measure (in correspondence with Ospina and Ferrari [2012]) as a square of the sample correlation coefficient between the response Y_t^c and the prediction of the conditional mean $\hat{\mu}_t^c$

$$R_2^2 = [\widehat{\text{corr}}(Y_t^c, \hat{\mu}_t^c)]^2 = \frac{\left[\sum_{t=1}^n (Y_t^c - \bar{Y}^c) (\hat{\mu}_t^c - \bar{\mu}^c) \right]^2}{\sum_{t=1}^n (Y_t^c - \bar{Y}^c)^2 \sum_{t=1}^n (\hat{\mu}_t^c - \bar{\mu}^c)^2},$$

where $\hat{\mu}_t^c = \hat{\alpha}_t c + (1 - \hat{\alpha}_t) \hat{\mu}_t$ with $\hat{\alpha}_t = h^{-1}(\mathbf{V}_t^\top \hat{\boldsymbol{\rho}})$, $\hat{\mu}_t = g^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\beta}})$ and $\bar{\mu}^c = \frac{1}{n} \sum_{t=1}^n \hat{\mu}_t^c$.

The residual analysis in the c -inflated beta regression model can be split into the analysis of the discrete part and the continuous part of the model. For the continuous part, we can use the standardised residuals described in Section 2.4, where we take only rows with $Y_t^c \in (0, 1)$

$$r_t^{c,C} = \frac{Y_t^c - \hat{\mu}_t^c}{\sqrt{\widehat{\text{var}}(Y_t^c | \mathbf{X}_t, \mathbf{V}_t, Y_t^c \in (0, 1))}}, \quad t \in \{1, \dots, n; Y_t^c \in (0, 1)\},$$

where $\widehat{\text{var}}(Y_t^c | \mathbf{X}_t, \mathbf{V}_t, Y_t^c \in (0, 1)) = \frac{\hat{\mu}_t(1 - \hat{\mu}_t)}{1 + \hat{\phi}}$.

For the discrete part, we can define the standardised residuals as

$$r_t^{c,D} = \frac{M_t - \hat{\alpha}_t}{\sqrt{\widehat{\text{var}}(M_t | \mathbf{X}_t, \mathbf{V}_t)}}, \quad t \in \{1, \dots, n\},$$

where $\widehat{\text{var}}(M_t | \mathbf{X}_t, \mathbf{V}_t) = \hat{\alpha}_t(1 - \hat{\alpha}_t)$. Notice that the defined residuals $r_t^{c,D}$ are the standardised residuals in the generalised linear model for binary data, where the response is $M_t \sim \text{Alt}(\alpha_t)$.

As in Section 2.4, plotting $r_t^{c,C}$ or $r_t^{c,D}$ against the index of observations, fitted values, or covariates should not reveal any unexpected patterns or extreme values.

More diagnostic tools can be found in Ospina and Ferrari [2012]. The authors propose adjusted Pearson residuals and randomised quantile residuals for the simulated envelope graph. In the article, influential observations are investigated using the likelihood displacements of each observation. There are also various information criteria for the model selection described.

3.3 Predictions

Since the simple approaches fit models to transformed data, we need to clarify what we are trying to model. For all cases, the aim is to model the conditional mean of the response. There are several options for prediction statistics; in this section, we propose two possibilities for each treatment (including the c -inflated

beta regression model). In the simulation study in Section 4.2 all predictions are compared.

Our suggested predictions come from the shape of the c -inflated beta distribution. We either predict the conditional mean for the c -inflated beta distribution or ignore the fact that the original data are c -inflated and predict the conditional mean estimate of the beta regression.

For the simple approaches, the prediction of μ_t , ignoring the fact that the data come from the c -inflated distribution, is as expected

$$\hat{\mu}_t = g^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\beta}}), \quad (3.10)$$

where $\hat{\boldsymbol{\beta}}$ are the regression coefficient estimates for a given model. For the prediction using the information that the data come from the c -inflated beta distribution, we can define using (3.4)

$$\hat{\mu}_t^c = \mathbf{E}[Y_t^c | \mathbf{X}_t] = \tilde{\alpha}c + (1 - \tilde{\alpha})\hat{\mu}_t = \tilde{\alpha}c + (1 - \tilde{\alpha})g^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\beta}}),$$

where we can estimate $\tilde{\alpha} = \frac{1}{n} \sum_{t=1}^n M_t$ – a simple estimate of a parameter in the alternative distribution without dependency on the regressors.

For the c -inflated beta regression, the prediction that ignores c -inflation is the same as for other models (3.10) using only the continuous part of the model. Note that it is exactly the same model (and therefore the predictions are also the same) as for the approach with omitting observations with $Y_t = c$. For the other case, we have the prediction mentioned in the confidence interval (3.9)

$$\hat{\mu}_t^c = \mathbf{E}[Y_t^c | \mathbf{X}_t] = \hat{\alpha}c + (1 - \hat{\alpha})\hat{\mu}_t = h^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\rho}})c + (1 - h^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\rho}}))g^{-1}(\mathbf{X}_t^\top \hat{\boldsymbol{\beta}}),$$

which is a straightforward estimate for the conditional mean of the c -inflated beta distribution (3.4), when the parameter α depends on the regressors.

There are other prediction options; for example, for the third simple treatment (multiplying all observations, Section 3.1.3), one can suggest transforming the predictions of Y_t^* back using the inverse formula of (3.1). This is equivalent to using the four-parameter beta distribution approach with $a = -\frac{\varepsilon}{2}$, $b = 1 + \frac{\varepsilon}{2}$ described in Section 2.1.5. However, this prediction brings a problem of predicting the original variable outside of the $(0, 1)$ interval, and therefore we do not consider this prediction and focus only on $\hat{\mu}_t$ and $\hat{\mu}_t^c$ in this work.

4. Simulations

In this chapter, we examine the methods proposed in Chapters 2 and 3 on the simulated data. We choose the parameters, generate covariates from selected distributions, and then generate the response given the chosen coefficients, covariates, and link function. The generation process will be described in more detail later.

Firstly, we investigate the global measures of fit proposed in 2.4 and propose recommendations for model selection for different link functions. Then, in Section 4.2, we investigate the performance of four possible solutions to c -inflated beta distribution described in Chapter 3. We compare the precision of the predictions in and out of the sample.

4.1 Model selection and global measures

In Section 2.4 we describe two global goodness-of-fit measures - R_1^2 and R_2^2 . Another measure can also be the deviance of the fitted model. In this simulation, we want to compare different models and, using these measures, identify the best-fitting model. We have defined the likelihood ratio, Wald and Rao tests for submodels in Section 2.3, which can be used in model building. In this simulation, we focus on models that differ by the chosen link function, and we want to find the best global goodness-of-fit measure for the selection of the link function.

4.1.1 Data generation

In this simulation we would like to compare beta regression models $\mathcal{M}_1, \dots, \mathcal{M}_5$ with logit, probit, complementary log-log, log-log and Cauchy link functions for the following simulation conditions

- sample size $n \in \{50, 200\}$,
- regression coefficients $\beta \in \left\{ \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix} \right\}$,
- dispersion coefficient $\phi \in \{5, 100\}$,
- the distribution of our regressor $X_{t2} \sim N(0, 1)$, $X_{t2} \sim t(2)$ or $X_{t2} \sim \text{Po}(2)$, where $N(0, 1)$ is the standard normal distribution, $t(2)$ is the Student's t -distribution with 2 degrees of freedom and $\text{Po}(2)$ is the Poisson distribution with parameter $\lambda = 2$,
- link function for data generation $g_0(\cdot) \in \{\text{logit}, \text{Cauchy}\}$.

In each simulation replication, we generate data that fulfil the beta regression model (Definition 3) by generating n independent samples of $\mathbf{X}_t = (1, X_{t2})^\top$ and $Y_t \sim B(g_0(\mathbf{X}_t^\top \beta), \phi)$.

Remark that due to the finite machine precision, some of the generated Y_t were rounded and represented in the memory as 0 or 1. Since such data would not fulfil the definition of a beta regression model from Chapter 2, we removed these observations from the sample. Therefore, our sample size is often lower than n (by single percentage units).

In this simulation, we want to examine the models deviance, R_1^2 , R_2^2 as defined in Section 2.4. The results are reported as two types of tables: the first one calculates averages of those statistics, the second shows a summary of which model achieves the best deviance, R_1^2 , R_2^2 in a given replication (as % of the total number of replications). This represents the real-life situation in which we have to choose which measure to use for a decision on the final chosen model. Tracking how many times a model is chosen for each measure shows how many times the measure was correct in deciding on the best model.

4.1.2 Simulation algorithm

For the chosen set of simulation parameters, repeat $N = 100$ times the following

1. Generate $\mathbb{X}$ and $\mathbf{Y}$.
2. Fit models $\mathcal{M}_1, \dots, \mathcal{M}_5$ as a beta regression model of $\mathbf{Y}$ given $\mathbb{X}$ with logit, probit, complementary log-log, log-log and Cauchy link functions. For this fit use default initial values of parameters (as suggested in Section 2.2.3).
3. If any of the models do not converge, try fitting the model with the vector $(\boldsymbol{\beta}^\top, \phi)^\top = (0, 0, 2)^\top$ as initial values.
4. Calculate deviance, R_1^2 , R_2^2 .
5. Find the model with lowest deviance, the model with highest R_1^2 and the model with highest R_2^2 .

After the repetitions, we calculate the convergence rate when using only original initial values and the convergence rate when non-convergent models are then initialised with $(0, 0, 2)^\top$, average of deviance, R_1^2 , R_2^2 and the relative number of decisions of a given model for each measure.

The reason why the algorithm does not converge with the initial values suggested in Section 2.2.3 is that extreme values are often passed to the optimisation function. If those extreme values are transformed or otherwise manipulated by an operation (e.g. calculating eigenvalues, matrix inversion), they can be rounded to non-suitable values for the algorithm (typically passing 0 to logarithm or singular Fisher information matrix).

We investigated those errors and our conclusion is that the case often occurs for extreme starting values of the parameter $\phi^{(0)}$. When $\phi^{(0)} > 10^9$, numerical problems appear repeatedly. Therefore, we suggest alternative initial values $(\boldsymbol{\beta}^\top, \phi) = (0, 0, 2)^\top$, leading to $\mathbf{X}_t \boldsymbol{\beta}^{(0)} = 0$ and

$$\mu^{(0)} = g^{-1}(\mathbf{X}_t \boldsymbol{\beta}^{(0)}) = \begin{cases} \frac{1}{2}, & \text{for logit, probit and Cauchy links,} \\ 1 - \frac{1}{e} \approx 0.632, & \text{for complementary log-log link,} \\ \frac{1}{e} \approx 0.368, & \text{for log-log link.} \end{cases} \quad (4.1)$$

For logit, probit, and Cauchy link, this choice represents starting the algorithm with the uniform distribution as stated in Section 1.1.3.

4.1.3 Results

The results of our simulations can be found in Tables 4.1, 4.2, 4.3, 4.4, 4.5, 4.6. The first two tables correspond to $X_{t2} \sim N(0, 1)$, the second two are generated using $X_{t2} \sim t(2)$, and the last two tables are for the covariate $X_{t2} \sim \text{Po}(2)$. The first table in each pair shows the mean of each of the measures examined, the second table shows the accuracy in the decision for the model (the decision is made by the lowest deviance, highest R_1^2 , and highest R_2^2 in each replication). In the first half of each table (logit link for data generation), the true model is the model $\mathcal{M}_1$ representing the model with the logit link; in the second half the true model is the model $\mathcal{M}_5$ fitted with the Cauchy link.

Detailed results

We will now comment on the results for each pair of tables corresponding to the given covariate distribution. Table 4.1 shows that for the logit link, the mean of R_2^2 is highest for model $\mathcal{M}_1$ except for the first set of parameters. Other measures show the best values for different models. For the Cauchy link, R_2^2 is the highest for model $\mathcal{M}_5$ in all cases. Notice that deviance often has the lowest values for model $\mathcal{M}_5$ regarding the link function used for data generation.

When we look at Table 4.2, we can see that for the selected measures, it was very difficult to select model $\mathcal{M}_1$ as the best model. For the logit link, the true model $\mathcal{M}_1$ was chosen the most often only in four cases, and all of them are for the measure R_2^2 . For the Cauchy link, we can see that both deviance and R_2^2 were very successful, but deviance was the lowest for model $\mathcal{M}_5$ in almost all cases, and also when the logit link was used for data generation.

For the case when the covariate follows Student's t -distribution, R_2^2 has the highest mean for the right model in all simulations (Table 4.3). R_2^2 also decided in all but one case the right model most often (see Table 4.4). We can see that for $\phi = 100$ and Cauchy link for data generation, all measures show satisfactory results. But for other sets of parameters, they are wrong in most of the replications.

For the Poisson distribution (Tables 4.5 and 4.6) we can see similar results as for the normal distribution. R_2^2 has the highest mean for the right model most of the time, but for the logit link, it chooses the wrong model the most often for many parameters. Other measures show unsatisfactory results when looking at both the mean and the accuracy of the decision. The only exception is the deviance and model with Cauchy link data generation, but as for other tables, that is mainly caused by the fact that the deviance was the lowest for model $\mathcal{M}_5$ regarding the link used for data generation.

When we compare the results with respect to the sample size n , we can see that all measures were more successful in selecting the right model for higher n with the exception of Cauchy link and R_1^2 . Looking at ϕ , we can see that, in general, the results were more accurate for higher ϕ , that is, when the distribution is more concentrated around its mean.

Summary of the results

Overall, we can see that only R_2^2 behaves as expected in most of the cases and identifies the true link for various parameters. The difference between the model

with logit, probit, complementary log-log, and log-log links is rather small, therefore, often even R_2^2 is not able to distinguish the right model. This also justifies the common usage of the logit link, because almost all other used alternatives (except the Cauchy link) provide similar results. R_2^2 also shows the highest values for data with a Cauchy link as the true link function. Therefore, our conclusion is the recommendation to use R_2^2 as a goodness-of-fit measure rather than R_2^1 (originally proposed in Ferrari and Cribari-Neto [2004]) or deviance of the model.

Discussion on convergence

We observe that for $X_{t2} \sim N(0, 1)$, the beta regression model works well, as expected. It is stable regarding the parameters choice, true link g_0 , and also the convergence rate, using the initial parameters suggested in Ferrari and Cribari-Neto [2004], is very high for all simulated experiments.

For the Student's t -distribution, we observed complications regarding the heavy tails. Because of the extreme values of the regressor, values of the response very close to boundaries can be generated, which leads to the problem described above of extremely high initial value for ϕ and non-convergence. On the other hand, the algorithm converged everytime for the suggested alternative initial parameters.

This case was also relevant to the regressor having a Poisson distribution, because the parameters were chosen such that the conditional mean of the response is closer to 1 for positive-valued distributions.

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	40.140	40.211	40.312	39.905	39.454
			R_1^2	0.254	0.255	0.251	0.245	0.162
			R_2^2	0.249	0.249	0.246	0.250	0.249
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 5$	200	D	157.209	157.618	158.433	155.960	154.092
			R_1^2	0.259	0.260	0.253	0.250	0.102
			R_2^2	0.248	0.247	0.246	0.247	0.246
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	87.230	86.320	84.577	87.117	84.373
			R_1^2	0.430	0.498	0.512	0.407	0.091
			R_2^2	0.480	0.479	0.472	0.476	0.459
			c. rate	100% (100%)	100% (100%)	99% (100%)	93% (100%)	36% (100%)
	$\phi = 5$	200	D	353.171	348.081	339.928	353.379	344.453
			R_1^2	0.412	0.496	0.512	0.387	0.029
			R_2^2	0.488	0.487	0.477	0.485	0.469
			c. rate	100% (100%)	100% (100%)	100% (100%)	77% (100%)	6% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	49.672	49.687	49.709	49.605	49.509
			R_1^2	0.850	0.851	0.848	0.844	0.819
			R_2^2	0.852	0.852	0.849	0.850	0.846
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	198.653	198.723	198.835	198.378	197.989
			R_1^2	0.850	0.851	0.846	0.844	0.810
			R_2^2	0.853	0.853	0.848	0.850	0.848
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
Cauchy	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	48.780	49.848	51.505	47.717	43.985
			R_1^2	0.908	0.918	0.905	0.896	0.534
			R_2^2	0.939	0.938	0.925	0.935	0.902
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	195.431	200.086	207.161	190.839	176.201
			R_1^2	0.911	0.921	0.904	0.896	0.391
			R_2^2	0.942	0.941	0.926	0.937	0.909
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	40.068	40.082	40.006	39.821	39.393
			R_1^2	0.300	0.304	0.303	0.286	0.178
			R_2^2	0.305	0.303	0.298	0.307	0.309
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 5$	200	D	156.491	156.539	156.484	155.364	154.072
			R_1^2	0.278	0.283	0.280	0.264	0.100
			R_2^2	0.286	0.285	0.280	0.286	0.288
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	48.920	48.181	47.187	48.854	48.190
			R_1^2	0.312	0.338	0.352	0.289	0.084
			R_2^2	0.396	0.387	0.368	0.405	0.423
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	92% (100%)
	$\phi = 5$	200	D	193.327	189.938	185.421	193.159	191.117
			R_1^2	0.326	0.358	0.375	0.300	0.030
			R_2^2	0.415	0.408	0.389	0.423	0.434
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	89% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	49.769	49.762	49.724	49.717	49.662
			R_1^2	0.853	0.852	0.842	0.852	0.837
			R_2^2	0.866	0.864	0.850	0.871	0.874
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	199.041	199.024	198.935	198.787	198.557
			R_1^2	0.848	0.848	0.838	0.846	0.829
			R_2^2	0.863	0.860	0.846	0.866	0.870
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	50.557	50.550	50.371	50.342	49.331
			R_1^2	0.817	0.811	0.779	0.825	0.757
			R_2^2	0.880	0.864	0.826	0.896	0.918
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	202.305	202.368	201.827	201.358	196.938
			R_1^2	0.826	0.818	0.784	0.833	0.744
			R_2^2	0.886	0.871	0.832	0.902	0.922
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)

Table 4.1: Model selection simulation with $X_{t2} \sim N(0, 1)$, summary for means of the measures

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	1%	0%	30%	7%	62%
			R_1^2	10%	20%	40%	18%	12%
			R_2^2	4%	3%	38%	19%	36%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 5$	200	D	0%	0%	10%	3%	87%
			R_1^2	26%	36%	24%	14%	0%
			R_2^2	6%	10%	30%	27%	27%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	2%	51%	0%	47%
			R_1^2	0%	37%	62%	1%	0%
			R_2^2	12%	19%	31%	25%	13%
			c. rate	100% (100%)	100% (100%)	99% (100%)	93% (100%)	36% (100%)
	$\phi = 5$	200	D	0%	0%	66%	0%	34%
			R_1^2	0%	21%	79%	0%	0%
			R_2^2	31%	24%	10%	31%	4%
			c. rate	100% (100%)	100% (100%)	100% (100%)	77% (100%)	6% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	6%	2%	92%
			R_1^2	23%	34%	32%	7%	4%
			R_2^2	23%	23%	25%	11%	18%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	0%	0%	100%
			R_1^2	33%	46%	16%	5%	0%
			R_2^2	36%	33%	8%	17%	6%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	0%	0%	0%	100%
			R_1^2	18%	67%	14%	1%	0%
			R_2^2	59%	28%	1%	12%	0%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	0%	0%	100%
			R_1^2	5%	94%	1%	0%	0%
			R_2^2	82%	14%	0%	4%	0%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
Cauchy	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	40%	10%	50%
			R_1^2	18%	18%	45%	12%	7%
			R_2^2	7%	4%	22%	21%	46%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 5$	200	D	0%	0%	19%	1%	80%
			R_1^2	14%	37%	45%	4%	0%
			R_2^2	15%	4%	13%	16%	52%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	2%	64%	0%	34%
			R_1^2	4%	18%	75%	1%	2%
			R_2^2	2%	2%	5%	16%	75%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	92% (100%)
	$\phi = 5$	200	D	0%	0%	81%	0%	19%
			R_1^2	1%	10%	89%	0%	0%
			R_2^2	7%	0%	0%	12%	81%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	89% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	28%	5%	67%
			R_1^2	19%	26%	13%	32%	10%
			R_2^2	9%	3%	0%	18%	70%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	7%	1%	92%
			R_1^2	45%	26%	2%	27%	0%
			R_2^2	1%	0%	0%	9%	90%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	0%	10%	0%	90%
			R_1^2	2%	9%	2%	80%	7%
			R_2^2	0%	0%	0%	4%	96%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	0%	0%	100%
			R_1^2	1%	1%	0%	98%	0%
			R_2^2	0%	0%	0%	0%	100%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)

Table 4.2: Model selection simulation with $X_{t2} \sim N(0, 1)$, summary for the relative accuracy of the decision on the best model

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	246.867	244.944	242.319	239.605	237.532
			R_1^2	0.587	0.622	0.568	0.509	0.354
			R_2^2	0.485	0.480	0.476	0.469	0.471
			c. rate	74% (100%)	76% (100%)	51% (100%)	51% (100%)	48% (100%)
		200	D	871.549	857.275	849.833	823.752	827.419
			R_1^2	0.515	0.630	0.458	0.454	0.263
			R_2^2	0.466	0.463	0.458	0.451	0.453
			c. rate	48% (100%)	45% (100%)	15% (100%)	12% (100%)	11% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	535.479	532.526	528.214	525.894	537.556
			R_1^2	0.588	0.705	0.594	0.450	0.288
			R_2^2	0.689	0.678	0.666	0.666	0.666
			c. rate	54% (100%)	54% (100%)	27% (100%)	18% (100%)	10% (100%)
		200	D	2085.498	2058.973	2037.578	2039.793	2084.784
			R_1^2	0.458	0.696	0.373	0.396	0.270
			R_2^2	0.696	0.689	0.672	0.660	0.672
			c. rate	14% (100%)	14% (100%)	1% (100%)	1% (100%)	0% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	162.173	161.261	163.377	161.059	149.322
			R_1^2	0.917	0.936	0.879	0.887	0.670
			R_2^2	0.936	0.933	0.925	0.917	0.916
			c. rate	79% (100%)	80% (100%)	68% (100%)	72% (100%)	74% (100%)
		200	D	572.425	566.985	543.922	560.829	470.241
			R_1^2	0.790	0.886	0.755	0.822	0.403
			R_2^2	0.937	0.936	0.927	0.929	0.917
			c. rate	65% (100%)	63% (100%)	33% (78%)	35% (99%)	55% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	256.799	257.996	220.045	238.735	231.915
			R_1^2	0.853	0.926	0.832	0.821	0.373
			R_2^2	0.975	0.971	0.952	0.945	0.947
			c. rate	79% (100%)	77% (100%)	35% (98%)	41% (99%)	56% (100%)
		200	D	1078.919	1072.553	865.827	1012.461	906.529
			R_1^2	0.575	0.812	0.559	0.677	0.240
			R_2^2	0.976	0.974	0.954	0.962	0.945
			c. rate	37% (100%)	31% (100%)	2% (51%)	0% (95%)	8% (100%)
Cauchy	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	43.280	42.679	39.469	41.895	45.244
			R_1^2	0.481	0.463	0.462	0.422	0.313
			R_2^2	0.457	0.444	0.425	0.440	0.482
			c. rate	96% (100%)	94% (100%)	82% (100%)	87% (100%)	90% (100%)
		200	D	173.228	162.338	154.300	152.878	179.848
			R_1^2	0.484	0.440	0.463	0.386	0.224
			R_2^2	0.448	0.432	0.415	0.408	0.467
			c. rate	90% (99%)	83% (100%)	37% (99%)	38% (99%)	69% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	66.469	65.431	62.886	64.935	68.611
			R_1^2	0.506	0.513	0.524	0.413	0.233
			R_2^2	0.591	0.553	0.521	0.562	0.651
			c. rate	89% (100%)	85% (100%)	55% (100%)	62% (100%)	64% (100%)
		200	D	240.259	234.515	236.266	245.375	264.172
			R_1^2	0.499	0.492	0.479	0.392	0.162
			R_2^2	0.610	0.571	0.532	0.559	0.649
			c. rate	77% (100%)	50% (100%)	5% (100%)	19% (100%)	24% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	54.548	57.376	53.788	56.802	49.629
			R_1^2	0.839	0.820	0.795	0.826	0.896
			R_2^2	0.912	0.899	0.856	0.890	0.936
			c. rate	99% (100%)	97% (99%)	88% (100%)	96% (100%)	98% (100%)
		200	D	218.701	237.079	233.313	247.514	198.310
			R_1^2	0.781	0.757	0.731	0.734	0.867
			R_2^2	0.922	0.914	0.849	0.884	0.935
			c. rate	92% (92%)	83% (86%)	59% (80%)	58% (97%)	96% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	60.124	62.403	55.565	62.727	49.000
			R_1^2	0.800	0.761	0.728	0.776	0.822
			R_2^2	0.929	0.891	0.828	0.892	0.966
			c. rate	96% (97%)	96% (99%)	79% (100%)	87% (99%)	96% (100%)
		200	D	244.273	269.969	237.853	263.521	196.429
			R_1^2	0.750	0.702	0.665	0.708	0.766
			R_2^2	0.930	0.901	0.833	0.875	0.968
			c. rate	83% (84%)	74% (83%)	30% (85%)	51% (93%)	96% (100%)

Table 4.3: Model selection simulation with $X_{t2} \sim t(2)$, summary for means of the measures

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	28%	24%	48%
			R_1^2	18%	44%	31%	6%	1%
			R_2^2	20%	10%	25%	21%	24%
			c. rate	74% (100%)	76% (100%)	51% (100%)	51% (100%)	48% (100%)
		200	D	0%	0%	25%	42%	33%
			R_1^2	16%	75%	3%	6%	0%
			R_2^2	55%	15%	11%	9%	10%
			c. rate	48% (100%)	45% (100%)	15% (100%)	12% (100%)	11% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	1%	13%	57%	23%	6%
			R_1^2	0%	66%	34%	0%	0%
			R_2^2	32%	14%	18%	20%	16%
			c. rate	54% (100%)	54% (100%)	27% (100%)	18% (100%)	10% (100%)
		200	D	0%	6%	51%	43%	0%
			R_1^2	0%	96%	4%	0%	0%
			R_2^2	91%	5%	0%	1%	3%
			c. rate	14% (100%)	14% (100%)	1% (100%)	1% (100%)	0% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	1%	2%	97%
			R_1^2	34%	41%	8%	16%	1%
			R_2^2	57%	29%	4%	9%	1%
			c. rate	79% (100%)	80% (100%)	68% (100%)	72% (100%)	74% (100%)
		200	D	0%	0%	0%	0%	100%
			R_1^2	33%	48%	1%	18%	0%
			R_2^2	79%	21%	0%	0%	0%
			c. rate	65% (100%)	63% (100%)	33% (78%)	35% (99%)	55% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	1%	1%	6%	92%
			R_1^2	14%	73%	5%	8%	0%
			R_2^2	87%	10%	0%	3%	0%
			c. rate	79% (100%)	77% (100%)	35% (98%)	41% (99%)	56% (100%)
		200	D	0%	0%	0%	0%	100%
			R_1^2	0%	81%	0%	19%	0%
			R_2^2	97%	3%	0%	0%	0%
			c. rate	37% (100%)	31% (100%)	2% (51%)	0% (95%)	8% (100%)
Cauchy	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	3%	1%	53%	29%	14%
			R_1^2	23%	12%	41%	12%	12%
			R_2^2	11%	2%	8%	15%	64%
			c. rate	96% (100%)	94% (100%)	82% (100%)	87% (100%)	90% (100%)
		200	D	0%	4%	46%	50%	0%
			R_1^2	47%	6%	39%	4%	4%
			R_2^2	6%	0%	0%	3%	91%
			c. rate	90% (99%)	83% (100%)	37% (99%)	38% (99%)	69% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	4%	13%	53%	20%	10%
			R_1^2	14%	21%	56%	4%	5%
			R_2^2	4%	0%	0%	6%	90%
			c. rate	89% (100%)	85% (100%)	55% (100%)	62% (100%)	64% (100%)
		200	D	14%	23%	47%	16%	0%
			R_1^2	24%	34%	36%	3%	3%
			R_2^2	1%	0%	0%	0%	99%
			c. rate	77% (100%)	50% (100%)	5% (100%)	19% (100%)	24% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	6%	7%	87%
			R_1^2	9%	0%	10%	15%	66%
			R_2^2	3%	0%	1%	6%	90%
			c. rate	99% (100%)	97% (99%)	88% (100%)	96% (100%)	98% (100%)
		200	D	0%	0%	6%	3%	91%
			R_1^2	14%	0%	4%	3%	79%
			R_2^2	0%	0%	0%	0%	100%
			c. rate	92% (92%)	83% (86%)	59% (80%)	58% (97%)	96% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	1%	1%	10%	8%	80%
			R_1^2	19%	0%	4%	24%	53%
			R_2^2	1%	0%	0%	0%	99%
			c. rate	96% (97%)	96% (99%)	79% (100%)	87% (99%)	96% (100%)
		200	D	1%	1%	8%	13%	77%
			R_1^2	22%	0%	5%	17%	56%
			R_2^2	0%	0%	0%	0%	100%
			c. rate	83% (84%)	74% (83%)	30% (85%)	51% (93%)	96% (100%)

Table 4.4: Model selection simulation with $X_{t2} \sim t(2)$, summary for the relative accuracy of the decision on the best model

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	61.554	61.050	60.294	61.653	60.098
			R_1^2	0.352	0.362	0.342	0.347	0.129
			R_2^2	0.285	0.284	0.281	0.285	0.275
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	63% (100%)
	$\phi = 5$	200	D	258.893	255.499	250.398	259.984	255.335
			R_1^2	0.344	0.370	0.347	0.337	0.045
			R_2^2	0.276	0.276	0.273	0.276	0.263
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	27% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	373.315	373.027	374.975	371.096	374.343
			R_1^2	0.389	0.439	0.442	0.386	0.071
			R_2^2	0.295	0.289	0.281	0.297	0.289
			c. rate	36% (100%)	41% (100%)	39% (100%)	35% (100%)	0% (100%)
	$\phi = 5$	200	D	1494.143	1501.113	1508.505	1493.574	1507.208
			R_1^2	0.370	0.419	0.421	0.369	0.029
			R_2^2	0.267	0.261	0.254	0.267	0.256
			c. rate	20% (100%)	19% (100%)	12% (100%)	22% (100%)	0% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	49.121	49.458	49.808	48.907	47.098
			R_1^2	0.863	0.864	0.855	0.860	0.654
			R_2^2	0.861	0.860	0.854	0.860	0.828
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	196.006	197.894	200.145	195.046	187.045
			R_1^2	0.862	0.867	0.856	0.857	0.472
			R_2^2	0.863	0.862	0.855	0.862	0.829
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
Cauchy	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	330.072	1146.672	79.644	76.487	59.252
			R_1^2	0.657	0.772	0.822	0.652	0.145
			R_2^2	0.873	0.866	0.846	0.873	0.815
			c. rate	88% (100%)	82% (100%)	69% (100%)	87% (100%)	16% (100%)
	$\phi = 100$	200	D	300.155	310.918	306.262	296.921	226.934
			R_1^2	0.598	0.755	0.817	0.592	0.044
			R_2^2	0.871	0.865	0.846	0.871	0.815
			c. rate	99% (100%)	94% (100%)	68% (100%)	99% (100%)	2% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	44.202	43.833	43.365	44.311	44.662
			R_1^2	0.206	0.207	0.199	0.204	0.095
			R_2^2	0.187	0.184	0.180	0.189	0.196
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	95% (100%)
	$\phi = 5$	200	D	171.999	170.499	168.423	172.450	173.672
			R_1^2	0.193	0.194	0.184	0.190	0.039
			R_2^2	0.168	0.166	0.162	0.169	0.173
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	94% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	108.040	108.090	108.227	108.014	108.738
			R_1^2	0.158	0.163	0.157	0.158	0.052
			R_2^2	0.118	0.116	0.114	0.119	0.128
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	31% (100%)
	$\phi = 5$	200	D	437.576	437.981	438.621	437.423	438.718
			R_1^2	0.150	0.154	0.145	0.149	0.025
			R_2^2	0.101	0.099	0.097	0.101	0.105
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	8% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	49.606	49.599	49.587	49.585	49.308
			R_1^2	0.727	0.723	0.713	0.730	0.695
			R_2^2	0.748	0.740	0.726	0.754	0.769
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	198.376	198.330	198.289	198.276	197.201
			R_1^2	0.719	0.714	0.703	0.721	0.674
			R_2^2	0.745	0.736	0.720	0.751	0.769
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	48.790	48.781	48.753	48.773	48.352
			R_1^2	0.588	0.590	0.588	0.587	0.485
			R_2^2	0.633	0.623	0.612	0.636	0.661
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	195.187	195.093	194.976	195.136	193.534
			R_1^2	0.577	0.577	0.574	0.576	0.461
			R_2^2	0.621	0.611	0.599	0.625	0.652
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)

Table 4.5: Model selection simulation with $X_{t2} \sim \text{Po}(2)$, summary for means of the measures

g_0	parameters	n	measure	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
logit	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	1%	2%	51%	0%	46%
			R_1^2	13%	45%	20%	20%	2%
			R_2^2	12%	4%	37%	19%	28%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	63% (100%)
	$\phi = 5$	200	D	0%	0%	82%	0%	18%
			R_1^2	11%	84%	3%	2%	0%
			R_2^2	16%	15%	24%	37%	8%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	27% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	20%	13%	2%	45%	20%
			R_1^2	6%	35%	56%	3%	0%
			R_2^2	14%	9%	21%	20%	36%
			c. rate	36% (100%)	41% (100%)	39% (100%)	35% (100%)	0% (100%)
	$\phi = 5$	200	D	11%	0%	0%	78%	11%
			R_1^2	0%	42%	58%	0%	0%
			R_2^2	14%	9%	5%	44%	28%
			c. rate	20% (100%)	19% (100%)	12% (100%)	22% (100%)	0% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	0%	0%	100%
			R_1^2	22%	44%	10%	24%	0%
			R_2^2	23%	13%	21%	42%	1%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	0%	0%	100%
			R_1^2	22%	68%	5%	5%	0%
			R_2^2	46%	27%	1%	26%	0%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	0%	7%	0%	93%
			R_1^2	0%	4%	96%	0%	0%
			R_2^2	27%	22%	1%	50%	0%
			c. rate	88% (100%)	82% (100%)	69% (100%)	87% (100%)	16% (100%)
	$\phi = 100$	200	D	0%	0%	0%	0%	100%
			R_1^2	0%	0%	100%	0%	0%
			R_2^2	55%	6%	0%	39%	0%
			c. rate	99% (100%)	94% (100%)	68% (100%)	99% (100%)	2% (100%)
Cauchy	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	2%	73%	1%	24%
			R_1^2	14%	20%	35%	24%	7%
			R_2^2	4%	5%	19%	11%	61%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	95% (100%)
	$\phi = 5$	200	D	0%	0%	84%	0%	16%
			R_1^2	27%	44%	9%	20%	0%
			R_2^2	10%	4%	3%	19%	64%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	94% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	3%	9%	36%	17%	35%
			R_1^2	8%	25%	35%	27%	5%
			R_2^2	1%	2%	29%	9%	59%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	31% (100%)
	$\phi = 5$	200	D	3%	11%	18%	35%	33%
			R_1^2	25%	34%	21%	20%	0%
			R_2^2	4%	1%	11%	23%	61%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	8% (100%)
	$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	50	D	0%	0%	26%	0%	74%
			R_1^2	9%	17%	7%	46%	21%
			R_2^2	4%	2%	0%	18%	76%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	5%	0%	95%
			R_1^2	7%	7%	0%	83%	3%
			R_2^2	0%	0%	0%	2%	98%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$	50	D	0%	0%	28%	0%	72%
			R_1^2	11%	14%	45%	22%	8%
			R_2^2	2%	3%	1%	18%	76%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)
	$\phi = 100$	200	D	0%	0%	12%	0%	88%
			R_1^2	30%	37%	14%	19%	0%
			R_2^2	0%	0%	0%	3%	97%
			c. rate	100% (100%)	100% (100%)	100% (100%)	100% (100%)	100% (100%)

Table 4.6: Model selection simulation with $X_{t2} \sim \text{Po}(2)$, summary for the relative accuracy of the decision on the best model

4.2 Comparison of inflated beta distribution treatments

In this simulation, we want to compare the suggested treatments of c -inflated beta distribution described in Chapter 3. We are interested in the prediction precision, and because some of the models are based on the adjusted response, we will mainly look at the prediction out of the training sample.

We are interested in 3 measures for our predictions - root mean square error ($RMSE$), mean absolute error (MAE) and mean bias (denoted simply as $bias$). They are defined in general in the following way

$$\begin{aligned} RMSE(\widehat{\mathbf{Z}}) &= \sqrt{\frac{1}{n} \sum_{t=1}^n (Z_t^0 - \widehat{Z}_t)^2}, \\ MAE(\widehat{\mathbf{Z}}) &= \frac{1}{n} \sum_{t=1}^n |Z_t^0 - \widehat{Z}_t|, \\ bias(\widehat{\mathbf{Z}}) &= \frac{1}{n} \sum_{t=1}^n (Z_t^0 - \widehat{Z}_t), \end{aligned}$$

where $\mathbf{Z}^0 = (Z_1^0, \dots, Z_n^0)^\top$ is a random vector, $\widehat{\mathbf{Z}} = (\widehat{Z}_1, \dots, \widehat{Z}_n)^\top$ is the prediction, and n is the size of the prediction sample (either train or test data). Since we will repeat the experiment N times, we will report the mean of those measures defined as

$$\begin{aligned} \overline{RMSE}(\widehat{\mathbf{Z}}) &= \frac{1}{N} \sum_{l=1}^N RMSE(\widehat{\mathbf{Z}}^{(l)}), \\ \overline{MAE}(\widehat{\mathbf{Z}}) &= \frac{1}{N} \sum_{l=1}^N MAE(\widehat{\mathbf{Z}}^{(l)}), \\ \overline{bias}(\widehat{\mathbf{Z}}) &= \frac{1}{N} \sum_{l=1}^N bias(\widehat{\mathbf{Z}}^{(l)}), \end{aligned}$$

where $\widehat{\mathbf{Z}}^{(l)}$ is the prediction in the l -th sample.

4.2.1 Data generation

For this simulation, we choose

- sample size $n \in \{50, 200\}$,
- regression coefficients $\boldsymbol{\beta} \in \left\{ \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix} \right\}$ and corresponding regression coefficients $\boldsymbol{\rho} \in \left\{ \begin{pmatrix} -2.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} -1.5 \\ 1.0 \end{pmatrix} \right\}$,
- dispersion coefficient $\phi \in \{5, 100\}$,
- the distribution of our regressor $X_{t2} \sim N(0, 1)$, $X_{t2} \sim t(2)$ or $X_{t2} \sim \text{Po}(2)$, where $N(0, 1)$ is the standard normal distribution, $t(2)$ is the Student's t -distribution with 2 degrees of freedom and $\text{Po}(2)$ is the Poisson distribution with parameter $\lambda = 2$.

For this simulation, we use the logit link function for both discrete and continuous models, so $g_0(\cdot) = h_0(\cdot) = \text{logit}$. In each iteration, we first generate the training data that satisfy the beta regression model (Definition 5) by generating n independent samples of $\mathbf{X}_t = (1, X_{t2})^\top$, $Y_t^c \sim \text{B}(g_0(\mathbf{X}_t^\top \boldsymbol{\beta}), \phi)$, and $M_t \sim \text{Alt}(h_0(\mathbf{X}_t^\top \boldsymbol{\rho}))$ (we take $\mathbf{V}_t = \mathbf{X}_t$), and then set $Y_t^c = c$ for $t : M_t = 1$. After that, we also generate test data of the same size n the same way as the training data.

Also, for this simulation, due to the finite machine precision, some of the generated Y_t^c were rounded and represented in the memory as 0 or 1 (before using M_t). Since such data would not fulfil the definition of a c -inflated beta regression model from Chapter 2 (the data do not follow a c -inflated beta distribution), we removed those observations from the sample. Therefore, our sample size is often lower than n (by single percentage units), and the sample size can be slightly different for the train and the test data, but thanks to the sufficient number of repetitions, the results are not affected.

4.2.2 Simulation algorithm

For chosen parameters of the simulation, logit link for all models, the choice $c = 1$ and $\varepsilon = 10^{-8}$ repeat $N = 300$ times the following

1. Generate $\mathbb{X}$ and $\mathbf{Y}^c$.
2. Fit model $\mathcal{M}_1$ as a beta regression model of $\mathbf{Y}^1$ given $\mathbb{X}^1$, where $\mathbf{Y}^1$ and $\mathbb{X}^1$ are created by omitting the rows where $Y_t^c = 1$.
3. Fit model $\mathcal{M}_2$ as a beta regression model of $\mathbf{Y}^2$ given $\mathbb{X}$, where $\mathbf{Y}^2$ is created by subtracting ε to observations where $Y_t^c = 1$.
4. Fit model $\mathcal{M}_3$ as a beta regression model of $\mathbf{Y}^3$ given $\mathbb{X}$, where $\mathbf{Y}^3$ is equal to transformed $\mathbf{Y}^c$ using the formula (3.1).
5. Fit model $\mathcal{M}_4$ as a c -inflated beta regression model of $\mathbf{Y}^c$ given $\mathbb{X}$. That means fitting the discrete model first using iterative weighted least squares and fitting the beta regression model for $\mathbf{Y}^1$ given $\mathbb{X}^1$.
6. If any of the models do not converge with the default initial values of parameters (suggested in Section 2.2.3), try fitting the model with the vector $(\boldsymbol{\beta}^\top, \phi)^\top = (0, 0, 2)^\top$ as initial values.
7. Calculate $RMSE(\widehat{\mathbf{Z}})$, $MAE(\widehat{\mathbf{Z}})$, $bias(\widehat{\mathbf{Z}})$ for predictions of models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$.

After the repetitions, we calculate the convergence rate when using only original initial values and convergence rate when non-convergent models are then initialised with $(0, 0, 2)^\top$, and averages of $\overline{RMSE}(\widehat{\mathbf{Z}})$, $\overline{MAE}(\widehat{\mathbf{Z}})$, $\overline{bias}(\widehat{\mathbf{Z}})$ for each of the models.

The reason for fitting the model with the alternative start parameters is described in Section 4.1 and because we use the logit link, the alternative initial values lead to a uniform distribution as stated in (4.1).

Since we are interested in the practitioners' point of view, we calculate all measures having $\mathbf{Z}^0 = \mathbf{Y}^c$, which is our response of the model. For all models, we compare all predictions described in Section 3.3.

4.2.3 Results

For all simulation scenarios, we calculate all measures for both $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^\top$ and $\hat{\boldsymbol{\mu}}^c = (\hat{\mu}_1^c, \dots, \hat{\mu}_n^c)^\top$ and compare them across the methods to find the best prediction of the response.

The results of our simulations can be found in Tables 4.7, 4.8, 4.9. The first table corresponds to $X_{t2} \sim N(0, 1)$, the second is generated using $X_{t2} \sim t(2)$, and the last table is for the covariate $X_{t2} \sim \text{Po}(2)$.

Because we found that for this simulation the model with an alternate start always converges, we present the convergence rate only for the first attempt to fit a model with a default choice of default parameters, which is often lower than 100% as discussed in Section 4.1.3.

Detailed results

The choice of parameter $\boldsymbol{\rho}$ corresponds to the expected behaviour in real applications; for a distribution with a mean around 0.5, the chance of observing 1 is lower than for extreme values (which explains the choice of ρ_1) and with increasing conditional mean because of the higher value of the regressor, the probability of observing 1 also increases (which explains the choice of ρ_2).

As expected, the performance of the models was better for train data than for test data for all measures in almost all simulation conditions. We can see that this difference is similar for all models, meaning that none of the models behaved unexpectedly when applied to the test data.

It is interesting that the measures *RMSE* and *MAE* were lower for distributions with heavier tails: $t(2)$ and $\text{Po}(2)$. This is mainly due to the fact that the probability density function of the conditional response on the left neighbourhood of 1 is lower than the probability of 1 (α_t). For the normal distribution, this disproportion is stronger than for the Student's t -distribution or the Poisson distribution. This difference decreases when we increase μ_t and therefore it shifts the continuous part of the distribution closer to 1.

Comparing the parameter sets, we conclude that for higher ϕ , all the measures are lower in most of the cases. That is mainly caused by the fact that with lower dispersion parameter, the chance of generating Y_t^c even lower than μ_t increases. The difference between prediction $\hat{\mu}_t$ and the response can be high, and for $\hat{\mu}_t^c$ even higher, because it is shifted towards 1 by the parameter $\alpha_t > 0$. Taking into account the choice of $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$, we can see that for the case with mean closer to 0.5 and lower α_t ($\boldsymbol{\beta} = (0.5, 0.5)^\top$, $\boldsymbol{\rho} = (-2.5, 0.5)^\top$) the results are worse than for the other case, where the conditional mean is usually closer to 1 and the probability of being 1 is higher. We assume that it is the same effect of disproportion of probability around 1 and directly in 1 as described in the previous paragraph.

Another result of our research is that the measures almost do not differ for different sample sizes n . Even for smaller samples, the model showed the same behaviour as for a four times larger sample size.

The most important observation of this simulation study is the choice of prediction ($\hat{\mu}_t$ or $\hat{\mu}_t^c$) and the model with the best results. Generally, we can see from the results of the test data that the measures *RMSE* and *MAE* are not very different in most cases, but the models $\mathcal{M}_2$ and $\mathcal{M}_3$ show slightly worse results than the models $\mathcal{M}_1$ and $\mathcal{M}_4$. For *bias* the difference between models $\mathcal{M}_2$ and

$\mathcal{M}_3$ and $\mathcal{M}_1$ and $\mathcal{M}_4$ is much higher.

When we look closer to the *RMSE* measure, we can see that in all simulations the best prediction is the prediction $\hat{\mu}_t^c$ of the model $\mathcal{M}_4$, the model of 1-inflated beta regression, where $\hat{\mu}_t^c$ is the estimate of the conditional mean. Notice that the second-best model is $\mathcal{M}_1$ fitted to data with omitted observations and its prediction $\hat{\mu}_t^c$. There are situations where prediction $\hat{\mu}_t$ of model $\mathcal{M}_1$ shows better results – the second set of parameters β, ρ for the $t(2)$ distribution.

With *MAE* the situation is similar, but the model $\mathcal{M}_4$ with prediction $\hat{\mu}_t^c$ is not the best in all cases. In about half of the conditions examined, another prediction has better *MAE*. However, in general $\hat{\mu}_t^c$ of $\mathcal{M}_4$ also provides a consistent performance for this measure, as it is always very close to the best prediction. It is worth noticing that for several experiments, also the models $\mathcal{M}_2$ and $\mathcal{M}_3$ show very similar results as $\mathcal{M}_4$. However, there are conditions in which these models are crucially worse.

Looking at the *bias*, we have to see the absolute dominance of prediction $\hat{\mu}_t^c$ from the model $\mathcal{M}_4$ – the bias is negligible in all cases. In some of the experiments, other models also approach zero bias, mostly $\mathcal{M}_1$ with the prediction $\hat{\mu}_t^c$ being the second best, but there are cases where $\mathcal{M}_2$ and $\mathcal{M}_3$ are also closer to zero bias. It is worth noticing that for all predictions of the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ have negative bias except $\hat{\mu}_t$ of $\mathcal{M}_1$. Negative bias means higher prediction than reality, which is caused by the transformation of the modelling dataset as discussed in Section 3.5. On the other hand, $\hat{\mu}_t$ of $\mathcal{M}_1$ brings positive bias, since we omitted the observations equal to 1 and therefore our prediction will often be lower than the response.

Summary of the results

Overall, $\mathcal{M}_4$ with prediction $\hat{\mu}_t^c$ provides consistent results across parameters, sample size, and regressor distributions. It is the best estimate in most of the cases, and if not, it is very close to the best prediction.

For $c = 0$ the problem is symmetrical and therefore similar results would be observed. We can conclude from our simulations that the recommendation is to use the prediction $\hat{\mu}_t^c$ of the model $\mathcal{M}_4$ when trying to predict the response in the c -inflated beta regression model.

parameters	n	measure	Train data				Test data			
			$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.207	0.218	0.218	0.207	0.214	0.226	0.227	0.214
		$\overline{RMSE}(\hat{\mu}^c)$	0.206	0.226	0.227	0.204	0.213	0.234	0.235	0.213
		$\overline{MAE}(\hat{\mu})$	0.171	0.176	0.176	0.171	0.177	0.182	0.183	0.177
		$\overline{MAE}(\hat{\mu}^c)$	0.170	0.181	0.182	0.168	0.175	0.188	0.188	0.174
		$bias(\hat{\mu})$	0.027	-0.058	-0.060	0.027	0.027	-0.059	-0.061	0.027
		$bias(\hat{\mu}^c)$	-0.004	-0.082	-0.083	0.000	-0.004	-0.082	-0.084	0.000
		c. rate	100%	100%	100%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.210	0.220	0.220	0.210	0.212	0.222	0.223	0.212
		$\overline{RMSE}(\hat{\mu}^c)$	0.209	0.229	0.229	0.208	0.211	0.231	0.231	0.210
		$\overline{MAE}(\hat{\mu})$	0.173	0.177	0.178	0.173	0.175	0.179	0.179	0.175
		$\overline{MAE}(\hat{\mu}^c)$	0.171	0.183	0.184	0.170	0.173	0.185	0.185	0.172
		$bias(\hat{\mu})$	0.028	-0.064	-0.065	0.028	0.028	-0.063	-0.065	0.028
		$bias(\hat{\mu}^c)$	-0.004	-0.088	-0.089	0.000	-0.004	-0.087	-0.089	0.000
		c. rate	100%	100%	100%	100%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.161	0.171	0.169	0.161	0.166	0.178	0.176	0.166
		$\overline{RMSE}(\hat{\mu}^c)$	0.164	0.186	0.186	0.158	0.169	0.193	0.192	0.163
		$\overline{MAE}(\hat{\mu})$	0.126	0.127	0.125	0.126	0.130	0.132	0.130	0.130
		$\overline{MAE}(\hat{\mu}^c)$	0.121	0.131	0.130	0.117	0.126	0.137	0.135	0.121
		$bias(\hat{\mu})$	0.030	-0.024	-0.028	0.030	0.027	-0.029	-0.033	0.027
		$bias(\hat{\mu}^c)$	-0.019	-0.061	-0.063	0.000	-0.023	-0.066	-0.069	-0.003
		c. rate	100%	93%	90%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.164	0.171	0.171	0.164	0.165	0.172	0.173	0.165
		$\overline{RMSE}(\hat{\mu}^c)$	0.166	0.188	0.189	0.161	0.167	0.189	0.190	0.162
		$\overline{MAE}(\hat{\mu})$	0.128	0.127	0.127	0.128	0.129	0.128	0.128	0.129
		$\overline{MAE}(\hat{\mu}^c)$	0.123	0.132	0.132	0.119	0.124	0.133	0.133	0.120
		$bias(\hat{\mu})$	0.029	-0.027	-0.028	0.029	0.030	-0.027	-0.028	0.030
		$bias(\hat{\mu}^c)$	-0.020	-0.063	-0.064	0.000	-0.020	-0.063	-0.064	0.000
		c. rate	100%	99%	99%	100%	-	-	-	-
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.107	0.133	0.135	0.107	0.107	0.143	0.144	0.107
		$\overline{RMSE}(\hat{\mu}^c)$	0.105	0.147	0.148	0.102	0.106	0.156	0.157	0.106
		$\overline{MAE}(\hat{\mu})$	0.062	0.115	0.116	0.062	0.063	0.120	0.121	0.063
		$\overline{MAE}(\hat{\mu}^c)$	0.068	0.130	0.132	0.066	0.070	0.136	0.137	0.068
		$bias(\hat{\mu})$	0.028	-0.070	-0.072	0.028	0.027	-0.069	-0.071	0.027
		$bias(\hat{\mu}^c)$	-0.004	-0.093	-0.094	0.000	-0.005	-0.092	-0.094	-0.001
		c. rate	100%	100%	100%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.110	0.134	0.135	0.110	0.110	0.136	0.138	0.110
		$\overline{RMSE}(\hat{\mu}^c)$	0.107	0.149	0.151	0.106	0.108	0.151	0.153	0.107
		$\overline{MAE}(\hat{\mu})$	0.062	0.116	0.117	0.062	0.063	0.117	0.119	0.063
		$\overline{MAE}(\hat{\mu}^c)$	0.068	0.135	0.136	0.066	0.068	0.136	0.138	0.066
		$bias(\hat{\mu})$	0.028	-0.076	-0.078	0.028	0.028	-0.075	-0.077	0.028
		$bias(\hat{\mu}^c)$	-0.004	-0.099	-0.100	0.000	-0.004	-0.099	-0.100	0.000
		c. rate	100%	100%	100%	100%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.084	0.100	0.101	0.084	0.085	0.105	0.106	0.085
		$\overline{RMSE}(\hat{\mu}^c)$	0.090	0.127	0.127	0.078	0.091	0.129	0.130	0.081
		$\overline{MAE}(\hat{\mu})$	0.053	0.080	0.080	0.053	0.054	0.082	0.082	0.054
		$\overline{MAE}(\hat{\mu}^c)$	0.068	0.101	0.102	0.053	0.068	0.102	0.103	0.055
		$bias(\hat{\mu})$	0.029	-0.035	-0.036	0.029	0.031	-0.033	-0.034	0.031
		$bias(\hat{\mu}^c)$	-0.020	-0.070	-0.070	0.000	-0.019	-0.068	-0.068	0.001
		c. rate	100%	96%	94%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.086	0.102	0.103	0.086	0.086	0.104	0.105	0.086
		$\overline{RMSE}(\hat{\mu}^c)$	0.090	0.130	0.131	0.080	0.091	0.131	0.132	0.081
		$\overline{MAE}(\hat{\mu})$	0.053	0.080	0.081	0.053	0.054	0.081	0.082	0.054
		$\overline{MAE}(\hat{\mu}^c)$	0.068	0.103	0.104	0.053	0.068	0.104	0.104	0.054
		$bias(\hat{\mu})$	0.029	-0.035	-0.036	0.029	0.029	-0.035	-0.036	0.029
		$bias(\hat{\mu}^c)$	-0.020	-0.069	-0.070	0.000	-0.019	-0.069	-0.070	0.000
		c. rate	100%	100%	99%	100%	-	-	-	-

Table 4.7: Simulation for comparison of 0-or-1 inflated beta distribution treatments with $X_{t2} \sim N(0, 1)$

parameters	n	measure	Train data				Test data			
			$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.199	0.211	0.213	0.199	0.207	0.220	0.222	0.207
		$\overline{RMSE}(\hat{\mu}^c)$	0.199	0.224	0.226	0.197	0.207	0.232	0.234	0.206
		$\overline{MAE}(\hat{\mu})$	0.161	0.170	0.171	0.161	0.168	0.176	0.178	0.168
		$\overline{MAE}(\hat{\mu}^c)$	0.162	0.179	0.181	0.158	0.169	0.185	0.188	0.165
		$\overline{bias}(\hat{\mu})$	0.027	-0.060	-0.063	0.027	0.027	-0.060	-0.063	0.027
		$\overline{bias}(\hat{\mu}^c)$	-0.012	-0.089	-0.092	-0.001	-0.012	-0.089	-0.092	-0.001
		c. rate	80%	78%	84%	80%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.203	0.214	0.216	0.203	0.205	0.215	0.217	0.205
		$\overline{RMSE}(\hat{\mu}^c)$	0.203	0.227	0.229	0.201	0.204	0.227	0.229	0.203
		$\overline{MAE}(\hat{\mu})$	0.164	0.172	0.174	0.164	0.166	0.172	0.174	0.166
		$\overline{MAE}(\hat{\mu}^c)$	0.165	0.181	0.183	0.161	0.166	0.181	0.183	0.162
		$\overline{bias}(\hat{\mu})$	0.026	-0.061	-0.065	0.026	0.029	-0.058	-0.062	0.029
		$\overline{bias}(\hat{\mu}^c)$	-0.012	-0.091	-0.094	0.000	-0.009	-0.088	-0.091	0.003
		c. rate	54%	50%	70%	54%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.156	0.166	0.175	0.156	0.162	0.171	0.181	0.162
		$\overline{RMSE}(\hat{\mu}^c)$	0.169	0.195	0.203	0.152	0.174	0.198	0.207	0.159
		$\overline{MAE}(\hat{\mu})$	0.119	0.124	0.132	0.119	0.124	0.128	0.137	0.124
		$\overline{MAE}(\hat{\mu}^c)$	0.126	0.138	0.145	0.110	0.130	0.141	0.149	0.116
		$\overline{bias}(\hat{\mu})$	0.027	-0.024	-0.026	0.027	0.027	-0.025	-0.026	0.027
		$\overline{bias}(\hat{\mu}^c)$	-0.033	-0.072	-0.073	-0.001	-0.034	-0.073	-0.074	-0.001
		c. rate	56%	36%	41%	56%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.159	0.168	0.174	0.159	0.163	0.172	0.178	0.163
		$\overline{RMSE}(\hat{\mu}^c)$	0.172	0.196	0.203	0.156	0.175	0.200	0.207	0.160
		$\overline{MAE}(\hat{\mu})$	0.122	0.126	0.131	0.122	0.125	0.129	0.133	0.125
		$\overline{MAE}(\hat{\mu}^c)$	0.128	0.140	0.144	0.113	0.131	0.143	0.148	0.116
		$\overline{bias}(\hat{\mu})$	0.027	-0.023	-0.029	0.027	0.026	-0.025	-0.031	0.026
		$\overline{bias}(\hat{\mu}^c)$	-0.033	-0.071	-0.076	0.000	-0.034	-0.073	-0.078	-0.001
		c. rate	19%	11%	18%	19%	-	-	-	-
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.103	0.132	0.135	0.103	0.105	0.141	0.143	0.105
		$\overline{RMSE}(\hat{\mu}^c)$	0.103	0.152	0.154	0.099	0.107	0.160	0.162	0.104
		$\overline{MAE}(\hat{\mu})$	0.058	0.112	0.114	0.058	0.060	0.117	0.119	0.060
		$\overline{MAE}(\hat{\mu}^c)$	0.069	0.133	0.135	0.061	0.071	0.137	0.139	0.064
		$\overline{bias}(\hat{\mu})$	0.026	-0.071	-0.073	0.026	0.027	-0.070	-0.072	0.027
		$\overline{bias}(\hat{\mu}^c)$	-0.012	-0.098	-0.100	0.000	-0.011	-0.097	-0.099	0.001
		c. rate	87%	84%	87%	87%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.106	0.139	0.141	0.106	0.107	0.141	0.143	0.107
		$\overline{RMSE}(\hat{\mu}^c)$	0.106	0.161	0.163	0.102	0.107	0.163	0.165	0.103
		$\overline{MAE}(\hat{\mu})$	0.059	0.118	0.120	0.059	0.060	0.120	0.122	0.060
		$\overline{MAE}(\hat{\mu}^c)$	0.070	0.142	0.144	0.062	0.071	0.144	0.146	0.063
		$\overline{bias}(\hat{\mu})$	0.027	-0.078	-0.080	0.027	0.027	-0.078	-0.081	0.027
		$\overline{bias}(\hat{\mu}^c)$	-0.013	-0.107	-0.109	0.000	-0.013	-0.107	-0.110	0.000
		c. rate	72%	69%	75%	72%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.080	0.113	0.116	0.080	0.079	0.118	0.122	0.079
		$\overline{RMSE}(\hat{\mu}^c)$	0.104	0.156	0.159	0.074	0.106	0.160	0.163	0.075
		$\overline{MAE}(\hat{\mu})$	0.048	0.086	0.088	0.048	0.049	0.089	0.091	0.049
		$\overline{MAE}(\hat{\mu}^c)$	0.077	0.117	0.119	0.048	0.078	0.120	0.122	0.049
		$\overline{bias}(\hat{\mu})$	0.026	-0.039	-0.042	0.026	0.026	-0.040	-0.042	0.026
		$\overline{bias}(\hat{\mu}^c)$	-0.036	-0.085	-0.087	0.000	-0.036	-0.086	-0.087	0.000
		c. rate	74%	45%	41%	74%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.082	0.119	0.123	0.082	0.082	0.121	0.126	0.082
		$\overline{RMSE}(\hat{\mu}^c)$	0.105	0.163	0.167	0.077	0.106	0.165	0.169	0.078
		$\overline{MAE}(\hat{\mu})$	0.048	0.090	0.092	0.048	0.049	0.091	0.094	0.049
		$\overline{MAE}(\hat{\mu}^c)$	0.077	0.121	0.123	0.048	0.078	0.122	0.125	0.049
		$\overline{bias}(\hat{\mu})$	0.026	-0.041	-0.043	0.026	0.026	-0.042	-0.045	0.026
		$\overline{bias}(\hat{\mu}^c)$	-0.036	-0.086	-0.088	0.000	-0.037	-0.088	-0.090	0.000
		c. rate	42%	19%	19%	42%	-	-	-	-

Table 4.8: Simulation for comparison of 0-or-1 inflated beta distribution treatments with $X_{t2} \sim t(2)$

parameters	n	measure	Train data				Test data			
			$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.163	0.166	0.167	0.163	0.168	0.169	0.169	0.168
		$\overline{RMSE}(\hat{\mu}^c)$	0.161	0.176	0.176	0.159	0.165	0.177	0.177	0.164
		$\overline{MAE}(\hat{\mu})$	0.131	0.125	0.125	0.131	0.135	0.128	0.128	0.135
		$\overline{MAE}(\hat{\mu}^c)$	0.123	0.126	0.126	0.122	0.127	0.128	0.128	0.126
		$\overline{bias}(\hat{\mu})$	0.031	-0.026	-0.027	0.031	0.033	-0.024	-0.025	0.033
		$\overline{bias}(\hat{\mu}^c)$	-0.010	-0.055	-0.055	0.000	-0.008	-0.053	-0.053	0.002
		c. rate	100%	89%	85%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.165	0.169	0.169	0.165	0.168	0.171	0.171	0.168
		$\overline{RMSE}(\hat{\mu}^c)$	0.164	0.179	0.179	0.162	0.166	0.181	0.181	0.165
		$\overline{MAE}(\hat{\mu})$	0.133	0.127	0.127	0.133	0.135	0.128	0.128	0.135
		$\overline{MAE}(\hat{\mu}^c)$	0.125	0.128	0.128	0.124	0.127	0.130	0.130	0.126
		$\overline{bias}(\hat{\mu})$	0.031	-0.027	-0.028	0.031	0.031	-0.028	-0.029	0.031
		$\overline{bias}(\hat{\mu}^c)$	-0.010	-0.057	-0.057	0.000	-0.011	-0.057	-0.058	-0.001
		c. rate	100%	94%	95%	100%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 5$	50	$\overline{RMSE}(\hat{\mu})$	0.084	0.086	0.087	0.084	0.089	0.087	0.087	0.089
		$\overline{RMSE}(\hat{\mu}^c)$	0.085	0.093	0.093	0.080	0.085	0.092	0.092	0.083
		$\overline{MAE}(\hat{\mu})$	0.058	0.052	0.052	0.058	0.060	0.053	0.053	0.060
		$\overline{MAE}(\hat{\mu}^c)$	0.044	0.044	0.044	0.044	0.045	0.045	0.045	0.046
		$\overline{bias}(\hat{\mu})$	0.023	-0.002	-0.002	0.023	0.024	-0.002	-0.002	0.024
		$\overline{bias}(\hat{\mu}^c)$	-0.012	-0.023	-0.023	0.000	-0.011	-0.023	-0.023	0.001
		c. rate	66%	1%	0%	66%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.086	0.092	0.095	0.086	0.088	0.094	0.097	0.088
		$\overline{RMSE}(\hat{\mu}^c)$	0.087	0.097	0.098	0.083	0.089	0.098	0.099	0.085
		$\overline{MAE}(\hat{\mu})$	0.058	0.055	0.058	0.058	0.058	0.056	0.059	0.058
		$\overline{MAE}(\hat{\mu}^c)$	0.044	0.046	0.047	0.045	0.045	0.047	0.048	0.045
		$\overline{bias}(\hat{\mu})$	0.022	0.001	0.003	0.022	0.021	0.000	0.003	0.021
		$\overline{bias}(\hat{\mu}^c)$	-0.012	-0.021	-0.020	0.000	-0.012	-0.022	-0.021	-0.001
		c. rate	76%	1%	1%	76%	-	-	-	-
$\beta = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.088	0.096	0.097	0.088	0.086	0.100	0.100	0.086
		$\overline{RMSE}(\hat{\mu}^c)$	0.085	0.113	0.114	0.081	0.085	0.116	0.117	0.081
		$\overline{MAE}(\hat{\mu})$	0.057	0.080	0.080	0.057	0.056	0.082	0.082	0.056
		$\overline{MAE}(\hat{\mu}^c)$	0.065	0.096	0.096	0.058	0.065	0.098	0.099	0.058
		$\overline{bias}(\hat{\mu})$	0.032	-0.033	-0.034	0.032	0.031	-0.034	-0.035	0.031
		$\overline{bias}(\hat{\mu}^c)$	-0.010	-0.061	-0.061	0.000	-0.011	-0.063	-0.063	-0.001
		c. rate	100%	94%	94%	100%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.087	0.096	0.097	0.087	0.087	0.098	0.098	0.087
		$\overline{RMSE}(\hat{\mu}^c)$	0.084	0.114	0.114	0.081	0.084	0.115	0.116	0.081
		$\overline{MAE}(\hat{\mu})$	0.056	0.079	0.080	0.056	0.056	0.080	0.081	0.056
		$\overline{MAE}(\hat{\mu}^c)$	0.064	0.097	0.097	0.057	0.064	0.098	0.098	0.057
		$\overline{bias}(\hat{\mu})$	0.031	-0.033	-0.034	0.031	0.031	-0.034	-0.035	0.031
		$\overline{bias}(\hat{\mu}^c)$	-0.010	-0.061	-0.062	0.000	-0.011	-0.062	-0.063	0.000
		c. rate	100%	99%	98%	100%	-	-	-	-
$\beta = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix}$ $\phi = 100$	50	$\overline{RMSE}(\hat{\mu})$	0.042	0.050	0.052	0.042	0.044	0.052	0.054	0.044
		$\overline{RMSE}(\hat{\mu}^c)$	0.046	0.060	0.060	0.035	0.047	0.060	0.061	0.039
		$\overline{MAE}(\hat{\mu})$	0.027	0.035	0.038	0.027	0.029	0.037	0.040	0.029
		$\overline{MAE}(\hat{\mu}^c)$	0.029	0.035	0.036	0.022	0.030	0.036	0.037	0.024
		$\overline{bias}(\hat{\mu})$	0.019	-0.003	-0.001	0.019	0.020	-0.004	-0.001	0.020
		$\overline{bias}(\hat{\mu}^c)$	-0.013	-0.022	-0.022	0.000	-0.013	-0.023	-0.022	0.000
		c. rate	99%	1%	1%	99%	-	-	-	-
	200	$\overline{RMSE}(\hat{\mu})$	0.044	0.055	0.055	0.044	0.044	0.055	0.055	0.044
		$\overline{RMSE}(\hat{\mu}^c)$	0.046	0.062	0.062	0.037	0.047	0.062	0.062	0.038
		$\overline{MAE}(\hat{\mu})$	0.028	0.040	0.040	0.028	0.028	0.040	0.040	0.028
		$\overline{MAE}(\hat{\mu}^c)$	0.029	0.037	0.037	0.023	0.030	0.037	0.037	0.023
		$\overline{bias}(\hat{\mu})$	0.019	-0.001	-0.001	0.019	0.019	-0.001	-0.001	0.019
		$\overline{bias}(\hat{\mu}^c)$	-0.013	-0.021	-0.021	0.000	-0.013	-0.022	-0.021	0.000
		c. rate	100%	1%	0%	100%	-	-	-	-

Table 4.9: Simulation for comparison of 0-or-1 inflated beta distribution treatments with $X_{t2} \sim \text{Po}(2)$

5. Applications on real data

In this chapter, we apply the beta regression model and the c -inflated beta regression model to real data, interpret the effects, and examine model diagnostics. In the first section, we examine a proportion of tertiary educated people in selected European countries. Our covariates are socioeconomic indicators, GDP, urbanisation, sex, and others.

The second illustration investigates data from a household survey in the Philippines. Since we want to show the application of the c -inflated beta regression model, we select proportion of household income spent on education as our response with other household characteristics as covariates.

5.1 Proportion of tertiary educated in Europe

Among the many applications mentioned in Section 1.4, an important part is socioeconomics and measuring indicators in certain groups of people. Such data are rarely available with single observations as they are more often made public and aggregated by groups.

5.1.1 Data

In this application, we investigate the proportion of tertiary educated people (people with college/university degree or higher education) in several European countries. The original data come from Eurostat and are measured each year. We took data from the year 2019, which are further divided by gender and density of population in the place where the person lives. We enriched this dataset by several socioeconomic indicators for each country. The data are available for EU28 (European Union members in 2019, including the United Kingdom), Iceland, North Macedonia, Norway, Serbia and Switzerland. Because the data are grouped by country, we do not consider the country as a covariate and examine the relationship between tertiary educated and other explanatory variables. In the dataset, we can find the following variables:

- Urbanisation: Density level in the area grouped into 3 categories: Cities, Rural areas, Towns and suburbs.
- Sex: Categorical with 2 levels: F and M.
- Tertiary: Proportion of people having tertiary education in a given country, urbanisation level, and sex group. Originally in %, therefore, divided by 100 for our model.
- GDP: The gross domestic product of a state in millions of Euros in 2019.
- Inhabitants: Number of inhabitants of a country in millions in 2019.
- Unemployment: Unemployment rate as % of the total population of a country (in %).
- Democracy: Country Democracy Index measured by The Economist Intelligence Unit (scale [0–10]).

- TOP_Universities: Number of universities in a country in TOP 10,000 universities in the world (in 2022). The ranking is done by Webometrics Ranking of World Universities, research group belonging to the largest public research body in Spain.
- GDP_per_capita: GDP per capita calculated as GDP/Inhabitants [in EUR].

The sources of each variable can be found in the *Application 1 - dataset.xlsx* attachment to this thesis.

The first 10 rows of our data set are printed below in Table 5.1.

Country	Urbanisation	Sex	Tertiary	GDP	Inhab.	Unemp.	Democracy	TOP_Uni.	GDP_p.c.
Austria	Cities	F	44.7	397518	8.859	3.3	8.29	15	44873
Austria	Cities	M	41.9	397518	8.859	3.3	8.29	15	44873
Austria	Rural areas	F	25.7	397518	8.859	3.3	8.29	15	44873
Austria	Rural areas	M	28.0	397518	8.859	3.3	8.29	15	44873
Austria	Towns and suburbs	F	30.9	397518	8.859	3.3	8.29	15	44873
Austria	Towns and suburbs	M	33.9	397518	8.859	3.3	8.29	15	44873
Belgium	Cities	F	45.2	478161	11.456	3.3	7.64	8	41741
Belgium	Cities	M	39.3	478161	11.456	3.3	7.64	8	41741
Belgium	Rural areas	F	45.5	478161	11.456	3.3	7.64	8	41741
Belgium	Rural areas	M	33.5	478161	11.456	3.3	7.64	8	41741

Table 5.1: The first 10 rows of the dataset for tertiary education modelling (some of the variable names are abbreviated)

5.1.2 Model fitting

First, we fitted the beta regression with all regressors (excluding Country_name and Country_code as reasoned above) and then excluded GDP having the p -value 0.33. In the next step, we fitted a model without the GDP covariate and excluded Unemployment having p -value 0.25. In the next model, TOP_Universities had p -value 0.16, so it was dropped. The model with Urbanisation, Sex, Inhabitants, Democracy and GDP_per_capita has all regression coefficients significant on significance level 0.05. Because of the satisfying results (high R^2_2) and the simplicity of the illustration, we do not consider interaction terms.

As the last step of this process, we verified that the model is not significantly worse than the first model with all coefficients. The likelihood ratio test with p -value 0.23 rejects the hypothesis of significant difference (at our confidence level 0.05).

The next step was to compare the link functions. Based on the simulation study in Section 4.1, we compare R^2_2 . The best R^2_2 was achieved with the log-log link function, for which $R^2_2 = 0.7093$.

Estimates of the regression coefficients can be found in Table 5.2 below.

Let us briefly comment on the coefficients. We can see that Urbanisation behaves as expected – proportion of tertiary educated is the highest for Cities, lower for Towns and suburbs, and even lower for Rural areas. Among women, the proportion of having university/college degree is higher than among men. Surprisingly, for more populated countries, the proportion decreases. Further data investigation confirmed that in less populated countries of the EU, the proportion of tertiary educated population is higher mainly in Cities and Towns and Suburbs. The Democracy Index follows an expected pattern, although there arises

	Estimate	Std. Error	z value	$\Pr(> z)$	
(Intercept)	−0.61167	0.1805	−3.3892	0.0007	***
Urbanisation – Rural areas	−0.56530	0.0366	−15.4383	<0.0001	***
Urbanisation – Towns and suburbs	−0.37263	0.0370	−10.0704	<0.0001	***
Sex – M	−0.21317	0.0295	−7.2358	<0.0001	***
Inhabitants	−0.00154	0.0007	−2.3060	0.0211	*
Democracy	0.10383	0.0268	3.8729	0.0001	***
GDP_per_capita	0.00466	0.0012	3.7540	0.0002	***
(phi)	38.66220	3.8448	10.0558	<0.0001	***
<i>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1</i>					

Table 5.2: Estimates for model of tertiary educated in chosen European countries

the question if the level of democracy increases university/college graduates or vice versa. A very significant predictor is also GDP_per_capita, and the data show that in the richer countries there are more tertiary educated.

5.1.3 Model diagnostics

The global measures of fit of our final model are

$$R_1^2 = 0.7001,$$

$$R_2^2 = 0.7093,$$

which indicates that our model could be a very good fit for our data. Let us examine the diagnostic graphs described in Section 2.4. They can be found in Appendix C.1.

In Figure C.1 we plot the Pearson and deviation residuals for the model. We cannot observe any detectable pattern, and none of the observations has an extreme residual. Any pattern in Figure C.2 could suggest a wrong choice of a link function as described in Section 2.4. Fortunately, there is no such pattern and therefore we can say that the model fits the data well.

Although the model fits the data very well (with an unexpectedly high R_2^2), there are still possible improvements to the model. One of them could be a further investigation of the Inhabitants variable, its transformation, or categorisation. Furthermore, GDP_per_capita is typically a candidate for the logarithmic transformation. The model can also be enriched with interaction terms.

5.2 Proportion of household income spent on education in the Philippines in 2015

In this case study, we would like to present a model of c -inflated beta regression to investigate the proportion of household income spent on education. We will assume (based on the nature of the data) that this variable follows a 0-inflated beta distribution, where 0 is observed for a household with no spending on education.

5.2.1 Data

The original dataset is acquired through the Kaggle dataset repository¹ and contains a sample of a Family Income and Expenditure Survey (FIES) conducted by the Philippine Statistics Authority (PSA). The survey is undertaken every three years, and its goal is to collect and analyse data of family income and expenditure given the characteristics of the family. More, including the final reports, can be found on the PSA website². The final report focusses on the distribution of income and expenditure throughout the country and examines the evolution over time.

The dataset we used is a sample of 2015 FIES, which covers income and expenditure in the year 2015. The original dataset has 60 variables and 41544 observations, where each observation represents the survey results of one household (also called family in the survey). The currency of the dataset is ₱ (Philippine peso) which is about 0.019 US dollar (April 2022).

Since our goal is to illustrate a c -inflated beta regression model, we take a sample of 1000 observations for model training and 1000 observations as testing data, which will later be used to compare the model with the simple treatments suggested in 3.1. We explore the variables and select 9 variables for our modelling purpose. The variables in our data set are:

- Household Type: factor variable describing the type of household income with 2 levels: Single Family/Nonfamily, Extended Family.
- Members: integer variable indicating the total number of family members.
- Members <5: integer variable indicating total number of family members younger than 5 years old.
- Members 5-17: integer variable indicating the total number of family members between 5 and 17 years old.
- Members Employed: integer variable indicating total number of family members employed.
- Head Age: age of the head of the family/household.
- Electricity: binary variable that indicates the availability of electricity for the household.
- Log Income: natural logarithm of total household income in 2015 [₱].
- Education Exp.: relative household education expenditure to household income in 2015.

The first 10 rows of our modelling dataset can be seen in Table 5.3.

5.2.2 Model fitting

Let us recall that the model of c -inflated beta regression can be fitted in two steps, the continuous model and the discrete model.

¹<https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure>

²<https://psa.gov.ph/content/family-income-and-expenditure-survey-fies-0>

Household Type	Members	Mem. <5	Mem. 5-17	Mem. employed	Head Age	Electricity	Log Income	Education Exp.
Single Family/Nonfamily	3	0	2	1	35	1	12.030	0.0069
Extended Family	6	1	3	2	39	1	13.336	0.1082
Single Family/Nonfamily	6	0	0	3	56	1	13.039	0.0000
Single Family/Nonfamily	4	1	1	1	24	0	11.294	0.0053
Single Family/Nonfamily	5	0	3	1	38	1	13.102	0.0898
Extended Family	8	0	3	1	47	1	12.046	0.0224
Single Family/Nonfamily	5	0	3	1	41	1	12.296	0.0926
Extended Family	6	0	2	1	77	1	12.200	0.0123
Extended Family	9	3	1	0	70	1	12.021	0.0006
Single Family/Nonfamily	8	1	4	0	47	1	12.288	0.0172

Table 5.3: The first 10 rows of the dataset for proportion of household income spending on education modelling

First, we fit the beta regression model of the response Education Exp. on all the other variables in the dataset with the addition of the square of Head Age, where observations with no education expenditures were omitted. This model used logit link function and has a relatively low $R^2_{2,c} = 0.0977$, as we would expect, due to a different purpose of the original study.

From this model, we exclude the non-significant regressor Electricity first and Household type in the second step, and find the best link function comparing $R^2_{2,c}$ as was suggested in Section 4.1. The highest $R^2_{2,c}$ was achieved for the Cauchy link function. The estimates of the final model can be found in Table 5.4.

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-40.1504	6.7218	-5.9731	<0.0001	***
Members	0.6332	0.2303	2.7497	0.0060	**
Members <5	-2.5837	0.6077	-4.2519	<0.0001	***
Members 5-17	-0.7537	0.2993	-2.5186	0.0118	*
Members Employed	-0.6526	0.2631	-2.4803	0.0131	*
Head Age	0.5356	0.1973	2.7143	0.0066	**
Log Income	1.5080	0.3744	4.0277	0.0001	***
(Head Age) ²	-0.0055	0.0019	-2.9126	0.0036	**
(phi)	19.1584	1.1877	16.1312	<0.0001	***

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

Table 5.4: Estimates of the model of proportion of household income spending on education in the Philippines – continuous part of the model

Looking at the coefficient estimates in Table 5.4, we can see that having more members in a family increases the proportion of income spent on education. On the other hand, with more children, the proportion decreases. This can imply that the main education expenditures of families are for college students. Colleges were paid in the Philippines until 2018, when the free tertiary education law³ in 2018 came into force. Surprisingly, for families with more members employed, the proportion is lower. One of the explanations could be that if a member is working, he/she does not go to school, and therefore the education expenditure proportion should be lower. We can see that age behaves as expected; it decreases the relative education expenditure for younger families and older ones (usually

³https://en.wikipedia.org/wiki/Universal_Access_to_Quality_Tertiary_Education_Act

elderly people), where typically there are no children in the household. The maximum is around 48 years old, which corresponds to the idea of children studying college or other higher education. A significant regressor is also the logarithm of household income; we can see from Table 5.3, that the values are around 12 and therefore the total effect of the covariate influences the expected relative education expenditures a lot. In particular, families with very low income spend also very low amounts on education.

The second part is the discrete model for $M_t = \mathbb{1}_0(Y_t^c)$. From that, we gradually exclude non-significant regressors Head Age (and leaving the square of Head Age in the model), Members <5, Electricity and choose Cauchy link as the link with the highest $R_{2,d}^2$. The estimates are shown in Table 5.5 below.

	Estimate	Std. Error	z value	$\Pr(> z)$	
(Intercept)	9.2594	2.4543	3.7727	0.0002	***
Household Type Single Family/Nonfamily	0.6035	0.3018	1.9997	0.0455	*
Members	-0.2665	0.1002	-2.6587	0.0078	**
Members 5-17	-4.1314	0.6036	-6.8449	<0.0001	***
Members Employed	0.3807	0.1482	2.5694	0.0102	*
Log Income	-0.7511	0.2026	-3.7067	0.0002	***
(Head Age) ²	0.0002	0.0001	2.5671	0.0103	*

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

Table 5.5: Estimates of the model of proportion of household income spending on education in the Philippines – discrete part of the model

When we look at the table of the estimates for the discrete model, we can see that it consists mostly of the same regressors as for the continuous model. In addition, there is a type of household in the discrete model, which suggests that individual families or nonfamily households are more likely not to spend on education. It makes sense that more members means a lower chance for not spending on education, because in larger families there is very often at least one student. The number of members between 5 to 17 years old is also significant, which contradicts the idea that there are expenditures only on college students. We can see that also primary and secondary education means expenditures for the family, even though from the continuous model we can see that higher expenditures comes from college students. We can have a similar explanation for the coefficient of employed members as for the continuous model. Each member can be either studying, working or retired, so with more employed members, less students in the family and therefore lower education expenditures. As in the continuous model, income is a very important regressor. Our model states that with higher income the chance of zero education spending is significantly lower. In the model, age is represented only by the square transformation, which strengthens the effect of families of elderly without education spending.

5.2.3 Model diagnostics

The global measures of fit of our final model are

$$\begin{aligned}R_{2,c}^2 &= 0.0991, \\ R_{2,d}^2 &= 0.5282, \\ R_2^2 &= 0.0801,\end{aligned}$$

where $R_{2,c}^2$, $R_{2,d}^2$ and R_2^2 are defined in Section 3.2.4. As we can see, numbers for the continuous part and the discrete part are very different. Having a high $R_{2,d}^2$ of the discrete model means that we are able to predict very well, if the observation is 0 or not. On the other hand, the prediction ability for the continuous part is not high, but as we show below in the diagnostic plots, it still provides a reasonable model.

From the diagnostic graphs in Appendix C.2 we can say that the model is not perfect; mainly it predicts higher education expenditure for the observations with the expenditure lower than 5% of the household income and the model also does not fit well to the observations with higher education expenditures. This can be caused by either missing predictors (wrong choice from the original dataset or missing transformations) or by the absence of regression dependency for this problem.

On the other hand, the fitted model is satisfactory for indicating whether the family has educational expenditures or not. For low relative education expenditure, the model is still useful, but the model defects appear for households with a higher proportion of income spent on education, where the predictions are much lower than the real values of household income spent on education.

5.2.4 Comparison of treatments

Similarly to Section 4.2, we would like to compare the treatments for our data. Using the same model formula and link, we also fitted the other three models for simple treatments described in Section 3.1. On test data (another sample of the original dataset with 1000 observations) we calculated $RMSE$, MAE , and $bias$ for all models for the predictions $\hat{\mu}$ and $\hat{\mu}^c$ described in Section 3.3.

The results are summarised in Table 5.6 below. As for the simulations, model $\mathcal{M}_1$ stands for model omitting observations, $\mathcal{M}_2$ for model adding $\varepsilon = 10^{-8}$, $\mathcal{M}_3$ for model multiplying each observation and $\mathcal{M}_4$ for the 0-inflated beta regression model described above.

In Table 5.6 we can see that the 0-inflated beta regression model ($\mathcal{M}_4$) provides the best predictions out of the training sample regarding $RMSE$. For MAE , the best model is the one with adding a small constant to the observations with education expenditure equal to 0. We can observe that all MAE values are lower than $RMSE$, which corresponds with the fact that all models predict well lower values of the response, but there are high prediction errors for observations with higher relative education expenditure. As in the simulation study in Section 4.2, $bias$ is much lower for the beta regression model and its prediction μ^c than for any other prediction.

Overall, comparing the four suggested treatments of c -inflated beta distribution, we came to the same conclusions as in the simulation study: the results are

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$RMSE(\hat{\mu})$	0.04808	0.04926	0.04958	0.04808
$RMSE(\hat{\mu}^c)$	0.04740	0.04920	0.04936	0.04684
$MAE(\hat{\mu})$	0.03399	0.02903	0.02907	0.03399
$MAE(\hat{\mu}^c)$	0.02855	0.02575	0.02578	0.02741
$bias(\hat{\mu})$	-0.00952	0.00058	0.00058	-0.00952
$bias(\hat{\mu}^c)$	0.00053	0.00766	0.00766	0.00010

Table 5.6: Comparison of four suggested treatments of c -inflated beta distribution applied on modelling proportion of household income spent on education in the Philippines in 2015

very similar, but c -inflated beta regression provides the best results regarding the root mean square error with the lowest bias from all observed possibilities.

Conclusion

The aim of this thesis was to describe the beta regression model to model data with a bounded interval range, to define extension to 0 or 1, and to perform a simulation study for various problems associated with the models. We have also shown two examples of beta regression applications, including estimates commentary and model diagnostics.

Firstly, we described the beta distribution and its characteristics. Special attention was paid to the shape of the probability density function. We summarised five possible shapes of the beta distribution density: constant (leading to uniform distribution), increasing, decreasing, “hill-shaped”, and “U-shaped”. For the last two options, we identified the mode of the distribution and the anti-mode of the distribution. The one-sided limits were calculated at 0 and 1. In this chapter, we presented the alternative parameterization used later in the beta regression model. Also, we described the four-parameter beta distribution for situations where the boundaries are not 0 and 1.

In the second chapter, we described the beta regression model as originally defined by Ferrari and Cribari-Neto [2004]. We described the connection to the generalised linear model, stated possible link functions, and compared them graphically. We enriched the work of Ferrari and Cribari-Neto with the interpretation of coefficients for the marginal change of a covariate and the ratio of the conditional means. The core of the theoretical part of the work is the extensive derivation of the parameter estimates, including the description of the algorithm. Unlike Ferrari and Cribari-Neto [2004], we derived the theory for stochastic regressors. We described the derivation steps in more detail, formalised and verified the assumptions of the maximum likelihood theory for the beta regression model, and added a rationale for the details of the individual steps. Statistical inference and model diagnostics were introduced as suggested in Ferrari and Cribari-Neto [2004] and Cribari-Neto and Zeileis [2010]. We created an illustrative example of diagnostics.

The third chapter was dedicated to the extension of the beta regression model, the c -inflated beta regression model, which can be used if the response takes one of the boundary values. We summarised three simple approaches often used by practitioners and a c -inflated beta regression model proposed by Ospina and Ferrari [2012]. We derived estimates in detail and connected them to the beta regression model. The two-step fitting algorithm is described and the equivalence of the first step to the generalised linear model fitting is explained. The second step corresponds to fitting the beta regression model using observations, where $Y_t^c \in (0, 1)$.

The practical part is a large part of the thesis. We conducted 2 simulations and built 2 beta regression models, including a c -inflated beta regression model. The aim of the first study was to determine the best global goodness-of-fit measure, which clearly indicates the right link function. The best results were observed for R_2^2 , although the difference between the links was negligible except for the Cauchy link function.

When producing the simulation data, we observed a phenomenon of a non-convergent fitting algorithm. Our investigation of the error led to the conclusion

that it was caused by an extremely high initial value of the dispersion parameter ϕ . Therefore, we proposed alternative initial values that correspond to the uniform distribution for logit, probit, and Cauchy links. For this alternative start, the model converged in almost all observed cases.

We have compared four proposed solutions for the c -inflated beta distribution, each with two different possible predictions. c -inflated beta regression model consistently provided the best results in all parameter choices. In most of the conditions studied, $RMSE$ and MAE were the lowest for the prediction based on both discrete and continuous parts of the c -inflated beta regression model, although the results of other treatments were similar. For *bias*, on the other hand, none of the other possible solutions is comparable to the c -inflated beta regression model, the predictions of which provide better results by order of magnitude.

In the last chapter, we applied the beta regression to real data. We analysed data of tertiary educated people, and education expenditure as a proportion of household income. All applications provided reasonable models with estimates that were in line with common sense. The diagnostics of the education expenditure model suggested a missing covariate or influential observations.

In the theoretical part, the thesis can be extended to beta regression with variable dispersion parameter ϕ as in Cribari-Neto and Zeileis [2010]. Another possible generalisation is possible by considering a general function of regressors and parameters instead of a linear function for the predictor. The theory using general predictor form is described in Ospina and Ferrari [2012]. Research of the beta regression model is still ongoing. Recently, an interesting article proposing two new parameter estimates that have better results on finite samples than the maximum likelihood estimate was written by Abonazel et al. [2022] among many others.

Bibliography

- A. H. Abdel-Karim. Extended Zero-one Inflated Beta and Adjusted Three-part Regression Models for Proportional Data Analysis. *Communications in Statistics – Simulation and Computation*, 46(8):6155–6172, 2017. doi: 10.1080/03610918.2016.1197248. URL <https://doi.org/10.1080/03610918.2016.1197248>.
- M. R. Abonazel, Z. Y. Algamal, F. A. Awwad, and I. M. Taha. A New Two-Parameter Estimator for Beta Regression Model: Method, Simulation, and Application. *Frontiers in Applied Mathematics and Statistics*, 7, 2022. doi: 10.3389/fams.2021.780322. URL <https://www.frontiersin.org/article/10.3389/fams.2021.780322>.
- M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, 1964. ISBN 978-0-486-61272-0.
- A. Bjorck. *Numerical methods for least square problems*. SIAM, 1996. ISBN 978-0-89871-360-2.
- F. Cribari-Neto and A. Zeileis. Beta Regression in R. *Journal of Statistical Software*, 34(2):1–24, 2010. doi: 10.18637/jss.v034.i02. URL <https://www.jstatsoft.org/index.php/jss/article/view/v034i02>.
- H. A. David. *Order Statistics*. 3rd Edition. Wiley, New Jersey, 2003. ISBN 0-471-38926-9.
- P. J. Davis. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Tenth Printing, New York, 1972. ISBN 978-0-486-61272-0.
- S. Ferrari and F. Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004. doi: 10.1080/0266476042000214501. URL <https://doi.org/10.1080/0266476042000214501>.
- J. W. Hardin and J. M. Hilbe. *Generalized Linear Model and Extensions*. Stata Press, 2007. ISBN 978-1-597-18225-6.
- N. L. Johnson. *Continuous Univariate Distributions Vol. 2*. 2nd Edition. Wiley, 1995. ISBN 978-0-471-58494-0.
- A. Jøsang. *A Formalism for Reasoning Under Uncertainty*. 1st ed. 2016 edition. Springer Verlag, 2016. ISBN 80-7378-001-1.
- E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Second Edition. Springer-Verlag, New York, 1998. ISBN 0-387-98502-6.
- P. MacCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989. ISBN 0-412-31760-5.

- W. K. Newey and D. McFadden. *Handbook of Econometrics, Vol.4*. Elsevier science, 1983. ISBN 978-0-444-88766-5.
- R. Ospina and S. L. P. Ferrari. Inflated Beta Distributions. *Statistical Papers*, 51(1):111–126, 2008. doi: 10.1007/s00362-008-0125-4. URL <https://doi.org/10.1007%2Fs00362-008-0125-4>.
- R. Ospina and S. L. P. Ferrari. A General Class of Zero-or-one Inflated Beta Regression Models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012. doi: <https://doi.org/10.1016/j.csda.2011.10.005>. URL <https://www.sciencedirect.com/science/article/pii/S0167947311003628>.
- L. Pace and A. Salvan. *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Advanced Series on Statistical Science & Applied Probability. World Scientific Publishing Company Pte Limited, 1997. ISBN 978-981-238-694-6.
- R. J. Rossi. *Mathematical Statistics : An Introduction to Likelihood Based Inference*. John Wiley & Sons, New York, 2018. ISBN 978-1-118-77104-4.

A. Gamma function

In this appendix, we recall the gamma function and its properties used in this work. More information on the gamma function, including proof of the properties of the function, can be found in Abramowitz and Stegun [1964].

The gamma function is defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

Notice from the definition that $\Gamma(1) = 1$. The function has a few properties that are used in this thesis, such as recurrence

$$\Gamma(z+1) = z\Gamma(z),$$

or the fact that $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

The functions derived from the gamma function used in the thesis are the digamma function $\psi(\cdot)$ and the trigamma function $\psi'(\cdot)$ defined as

$$\begin{aligned}\psi(z) &= \frac{\partial \log \Gamma(z)}{\partial z} = \frac{\frac{\partial}{\partial z} \Gamma(z)}{\Gamma(z)}, \\ \psi'(z) &= \frac{\partial \psi(z)}{\partial z} = \frac{\partial^2 \log \Gamma(z)}{\partial z^2}.\end{aligned}$$

The gamma function is also related to the beta function

$$\mathcal{B}(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt,$$

in the following way

$$\mathcal{B}(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},$$

proof can be found in Davis [1972].

B. Regularity conditions

Let us have random sample $Z = (Z_1, \dots, Z_n)$ of independent identically distributed random variables (or vectors), each with density $f(z|\boldsymbol{\theta}_Z)$ with respect to a σ -finite measure μ . We assume that $f(z|\boldsymbol{\theta}_Z) \in \mathcal{F}$, where

$$\mathcal{F} = \{\text{distributions with density } f(z|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq R^d\}$$

represents a parametric model for the distribution of the data. The regularity conditions are as follows:

- (R1) The number of parameters d in the model is constant.
- (R2) The support set $S = \{z \in \mathbb{R} : f(z|\boldsymbol{\theta}) > 0\}$ does not depend on the parameter $\boldsymbol{\theta}$.
- (R3) The parameter space Θ is an open set.
- (R4) The density $f(z|\boldsymbol{\theta})$ is a sufficiently smooth function of $\boldsymbol{\theta}$ (at least twice continuously differentiable).
- (R5) The Fisher information matrix $I(\boldsymbol{\theta})$ is finite, regular, and positive definite in a neighbourhood of $\boldsymbol{\theta}_Z$.
- (R6) The order of differentiation and integration can be interchanged in expressions such as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int h(z, \boldsymbol{\theta}) d\mu(z) = \int \frac{\partial}{\partial \boldsymbol{\theta}} h(z, \boldsymbol{\theta}) d\mu(z),$$

where $h(z, \boldsymbol{\theta})$ is either $f(z|\boldsymbol{\theta})$ or $\frac{\partial}{\partial \boldsymbol{\theta}} f(z, \boldsymbol{\theta})$.

C. Model diagnostic graphs

C.1 Proportion of tertiary educated in Europe

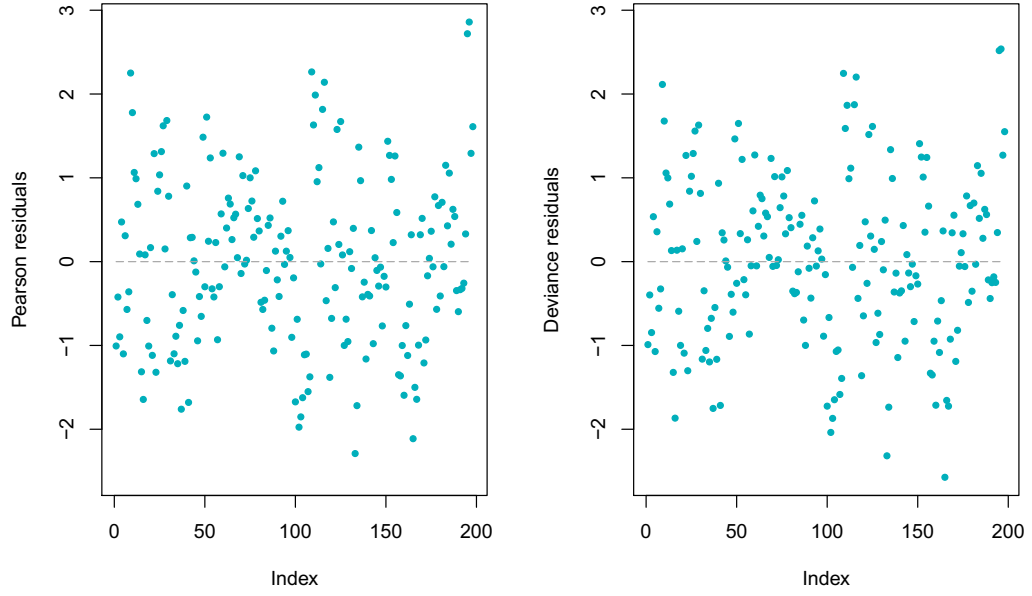


Figure C.1: Pearson and deviance residuals for the model of tertiary education

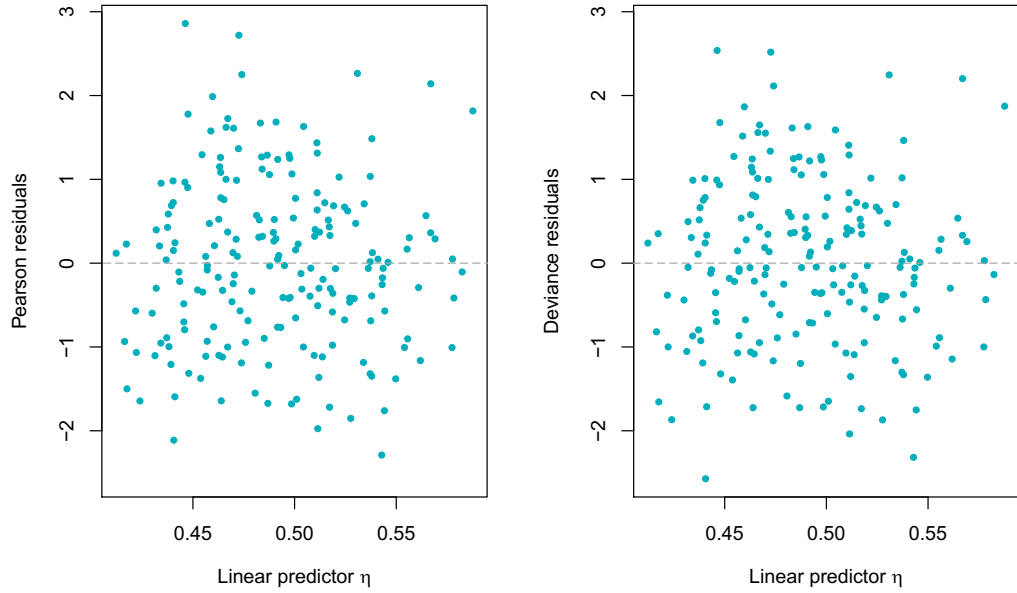


Figure C.2: Pearson and deviance residuals against the linear predictor η for the model of tertiary education

C.2 Proportion of household income spent on education in the Philippines in 2015

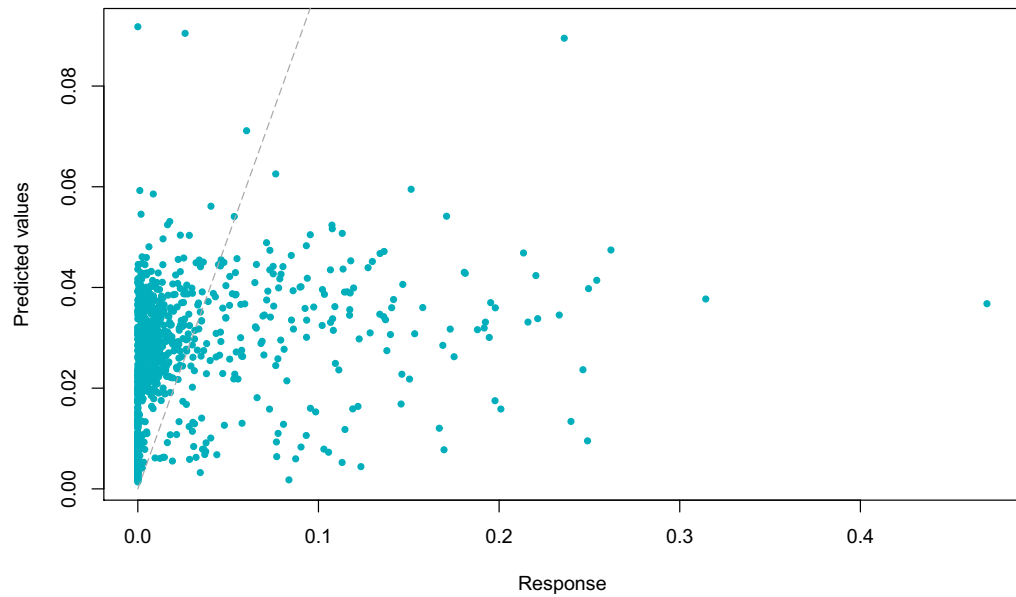


Figure C.3: Predictions against the response for the model of proportion of household income spending on education in the Philippines

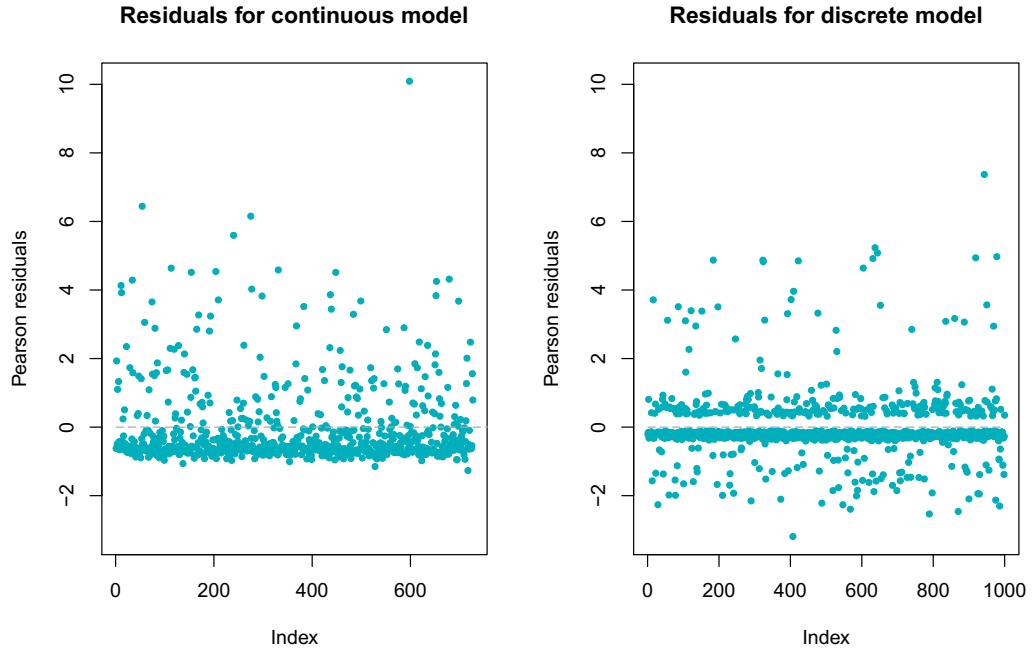


Figure C.4: Pearson residuals against the index for the continuous and discrete part of the model of proportion of household income spending on education in the Philippines

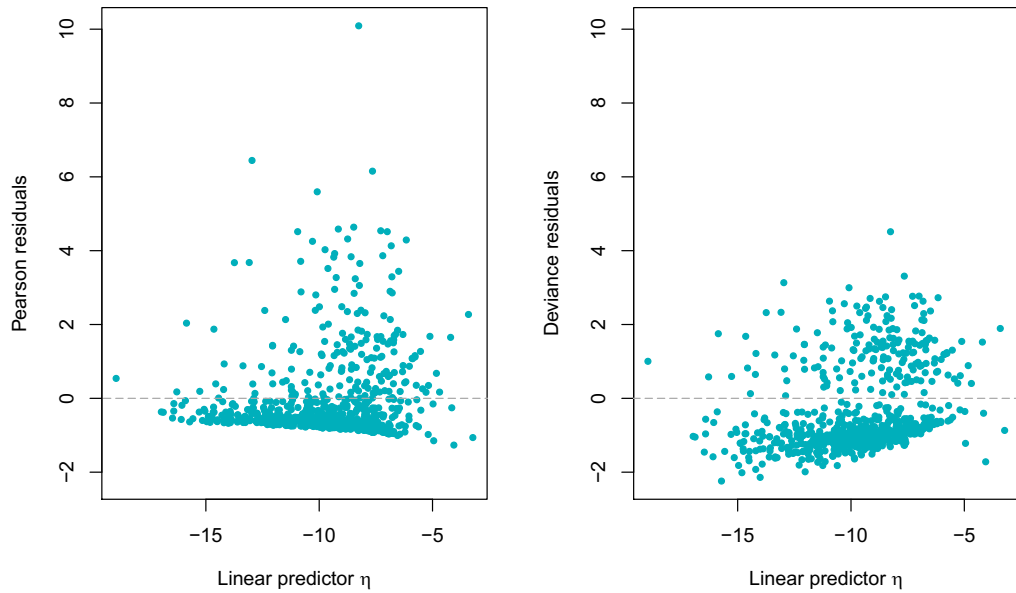


Figure C.5: Pearson and deviance residuals against the linear predictor η for the continuous part of the model of proportion of household income spending on education in the Philippines