Xacur *Journal of Statistical Distributions and Applications* (2019) 6:13
https://doi.org/10.1186/s40488-019-0102-6

Journal of Statistical Distributions and Applications

RESEARCH Open Access

The unifed distribution

Oscar Alberto Quijano Xacur

Correspondence:
oscar.quijano@use.startmail.com
Concordia University, Montreal, Canada

Abstract

We introduce a new distribution with support on (0,1) called unifed. It can be used as the response distribution for a GLM and it is suitable for data aggregation. We make a comparison to the beta regression. A link to an R package for working with the unifed is provided.

Keywords: Exponential dispersion family, GLM, R, Beta regression

Introduction

We introduce the unifed distribution. It is a continuous distribution with support on the interval (0,1). It can be characterized as the only exponential dispersion family containing the uniform distribution. This makes it suitable to be used as the response variable of a Generalized Linear Model (GLM).

An R (see (R Core Team 2017) and (Quijano Xacur 2019b)) package has been developed to work with this distribution. It is called `unifed` and contains functions for the density, distribution, quantiles and random generator. It also contains a family that can be used within the `glm` function of R. Additionally, the package provides Stan (Stan Development Team 2018) code for performing Bayesian analysis with the unifed including a function for fitting Bayesian unifed GLMs. Information about the package and how to install it can be found at https://gitlab.com/oquijano/unifed.

This is not the only model for performing regression on the unit interval. The beta regression (see (Ferrari and Cribari-Neto 2004)) has existed for a while and it provides more flexible shapes than the unifed GLM. One appealing property of the unifed GLM is that it is suitable for data reduction while the beta regression is not. This is discussed in "On the difficulties of data aggregation for the beta regression" section.

This paper is divided into 4 sections. In "Exponential dispersion families and GLMs" section we review the definition and properties of exponential dispersion families and GLMs. "The unifed distribution" section defines the unifed distribution. In "An illustrative example" section we illustrate an application to an auto insurance claims example. "Comparison between the unifed GLM and the beta regression" section reviews the beta regression and underlines it's differences with the unifed GLM.

Exponential dispersion families and GLMs

A reproductive Exponential Dispersion Family (EDF) is a set of distributions whose densities are given by

$$f(y|\theta,\phi) = a(y,\phi)\exp\left(\frac{1}{\phi}\left\{y\theta - \kappa(\theta)\right\}\right), \qquad \theta \in \Theta, \phi \in \Phi. \tag{1}$$

θ and Θ are called the canonical parameter and canonical space, respectively and ϕ is known as the dispersion parameter. For $\theta \in \text{int}(\Theta)$ (here int stands for interior),

$$\mathbb{E}[Y] = \dot{\kappa}(\theta) \qquad \text{and} \qquad \mathbb{V}[Y] = \phi\ddot{\kappa}(\theta), \tag{2}$$

where $\dot{\kappa} = \kappa'$ and $\ddot{\kappa} = \dot{\kappa}'$. (Eq. 2) allows to relate the mean and the variance and the mean of any EDF. This motivates the following definitions (see (Jørgensen 1997) or (Jørgensen 1992)).

Definition 0.1. *Given an exponential dispersion family, the mean domain of the family is defined as*

$$\Omega = \{\mu = \dot{\kappa}(\theta) : \theta \in int(\Theta)\}.$$

Definition 0.2. *The variance function of an EDF is defined as* $V : \Omega \to [0,\infty)$ *with*

$$V(\mu) = (\dot{\kappa} \circ \dot{\kappa}^{-1})(\mu).$$

Note that $\mathbb{V}[Y] = \phi\mathbf{V}(\mu)$. The support of the members of an EDF depend only on ϕ (and not on θ). For a given family, let C_ϕ be the convex support of any member of the family with dispersion parameter ϕ. We define the convex support of the family as

$$C_\Phi = \bigcup_{\phi\in\Phi} C_\phi.$$

Definition 0.3. *The unit deviance function of an exponential dispersion family is defined as* $d : C_\Phi \times \Omega \to [0,\infty)$ *with*

$$d(y,\mu) = 2\left[\sup_{\theta\in\Theta}\{\theta y - \kappa(\theta)\} - y\dot{\kappa}^{-1}(\mu) + \kappa\left(\dot{\kappa}^{-1}(\mu)\right)\right]. \tag{3}$$

The unit deviance function allows to re-parametrize (Eq. 1) as

$$f(y|\mu,\phi) = c(y,\phi)\exp\left(-\frac{1}{2\phi}d(y,\mu)\right). \tag{4}$$

This is known as the mean–value parametrization. When the canonical space Θ is open, the EDF is said to be regular. In this case $C_\Phi = \Omega$ and (Eq. 3) is equivalent to

$$d(y,\mu) = 2\left[y\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\} - \kappa\left(\dot{\kappa}^{-1}(y)\right) + \kappa\left(\dot{\kappa}^{-1}(\mu)\right)\right]. \tag{5}$$

Weights and data aggregation

In many applications it is useful to include a known positive weight to each observation. When this is done, the dispersion parameter is divided by the weight w, and (Eq. 1) and (Eq. 4) become respectively

$$f(y|\theta,\phi) = a(y,\phi/w)\exp\left(\frac{w}{\phi}\left\{y\theta - \kappa(\theta)\right\}\right), \quad \text{and}$$
$$f(y|\mu,\phi) = c(y,\phi/w)\exp\left(-\frac{w}{2\phi}d(y,\mu)\right). \tag{6}$$

There is a useful property of reproductive exponential dispersion families that allows for data aggregation. Jørgensen's notation (from (Jørgensen 1997)) is very convenient to express this property: given a fixed exponential family, if Y has mean μ and density given

by (Eq. 6), we say that it is $ED(\mu, \phi/w)$ distributed. The property is then as follows: if $Y_1, Y_2, \cdots, Y_n$ are independent, and $Y_i \sim ED(\mu, \phi/w_i)$, then

$$\bar{Y} = \frac{w_1 Y_1 + \cdots + w_n Y_n}{w_+} \sim ED(\mu, \phi/w_+), \qquad w_+ = \sum_{i=1}^{n} w_i. \tag{7}$$

GLMs

In a GLM the response variable is assumed to follow an EDF with density

$$f(y|\theta, \phi) = a(y, \phi) \exp\left(\frac{w}{\phi}\{y\theta - \kappa(\theta)\}\right). \tag{8}$$

Note that ϕ in (Eq. 1) corresponds to ϕ/w in (Eq. 8) which implies that the mean and variance can be expressed as $\mu = \kappa'(\theta)$ and $\sigma^2 = \phi\kappa''(\theta)/w$, respectively. Here $w \geq 0$ is known as the weight. In applications w is usually known and ϕ needs to be estimated. It is further assumed that there is a vector of explanatory variables, also known as covariates, $x = (x_1 \cdots x_p)^T$, a vector of coefficients $\mathcal{B} = (\mathcal{B}_0\ \mathcal{B}_1 \cdots \mathcal{B}_p)^T$ and a function g known as the link function such that

$$g(\mu) = \mathcal{B}_0 + x_1\mathcal{B}_1 + \cdots + x_p\mathcal{B}_p. \tag{9}$$

It is useful for further developments to express the canonical parameter θ in terms of the coefficients. Since $\mu = \kappa'(\theta) \equiv \dot{\kappa}(\theta)$ then:

$$\begin{aligned}(g \circ \dot{\kappa})(\theta) &= \mathcal{B}_0 + x_1\mathcal{B}_1 + \cdots + x_p\mathcal{B}_p \\ \theta &= (g \circ \dot{\kappa})^{-1}(\mathcal{B}_0 + x_1\mathcal{B}_1 + \cdots + x_p\mathcal{B}_p).\end{aligned} \tag{10}$$

The population can be divided into different classes according to the values of the explanatory variables. Thus, given a sample, we can group together all the observations that share the same values of the explanatory variables and aggregate them using (Eq. 7). It is important to mention that with this grouping there is no loss of information for estimating the mean since $\bar{Y}$ is a sufficient statistic for θ (but not for ϕ, thus some information is lost for the estimation of ϕ). In this sense we say that GLMs are suitable for *data aggregation*. At the end of "An illustrative example" section we illustrate this property with real data for a unifed GLM.

Possibly after aggregating, let m be the number of classes and $\theta \in \Theta^m$, where $\Theta^m = \left\{\theta = (\theta_1 \cdots \theta_m)^T : \theta_1, \ldots, \theta_m \in \Theta\right\}$ is the set of all possible values of the vector θ. The density of the sample can be expressed as

$$f(y|\theta, \phi) = A(y, \phi) \exp\left(\frac{y^T W\theta - 1^T W\kappa(\theta)}{\phi}\right), \qquad y \in \mathbb{R}^m, \tag{11}$$

where $\kappa(\theta) = \left(\kappa(\theta_1) \cdots \kappa(\theta_m)\right)^T$, $W = \text{diag}(w_1, \cdots, w_m)$, with w_i being the sum of all the weights in the i-th class, $1 = (1 \cdots 1)^T$ and $A(y, \phi) = \prod_{i=1}^{m} \left(a(y_i, \frac{w_i}{\phi})\right)$.

It is useful to reparameterize (Eq. 11) in terms of the mean vector μ instead of θ. Using the mean value parametrization (this is (Eq. 4) but substituting ϕ for ϕ/w), (Eq. 11) can be reparameterized as

$$f(y|\mu, \phi) = C(y, \phi) \exp\left(-\frac{1}{2\phi} D(y, \mu)\right), \tag{12}$$

where $C(y, \phi) = \prod_{i=1}^{m} c(y_i, \frac{\phi}{w_i})$, and $D : \Omega^m \times \Omega^m \to [0, \infty)$ with

$$D(y,\mu) = \sum_{i=1}^{m} w_i d(y_i,\mu_i), \tag{13}$$

$\Omega^m = \left\{(\mu_1 \cdots \mu_m)^T : \mu_1,\ldots,\mu_m \in \Omega\right\}$. D is called the deviance of the model. Note that finding the maximum likelihood estimator of $\mathcal{B}$ is equivalent to finding what value of $\mathcal{B}$ minimizes the deviance. For further details about the use and properties of the deviance see (Jørgensen 1992).

The unifed distribution

The unifed family is the Exponential Dispersion Family (EDF) generated by the uniform distribution (see Chapters 2 and 3 of (Jørgensen 1997) to see how an EDF can be generated from a moment generating function). We created the R package `unifed` (see (Quijano Xacur 2019b)) that includes functions to work with the unifed. In this section we make references to some functions in the package and we use `this font format` for those references.

To express the density of the unifed distribution we need the density of the sum of n independent $uniform(0,1)$ random variables. This corresponds to the Irwin-Hall distribution (see (Johnson et al. 1995)) and its density function is

$$h(y;n) = \frac{1}{(n-1)!}\sum_{k=0}^{\lfloor y\rfloor}(-1)^k\binom{n}{k}(y-k)^{n-1}, \qquad y\in[0,n], n\in\mathbb{N}. \tag{14}$$

The canonical and index spaces of the unifed family are $\Theta = \mathbb{R}$ and $\Phi = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\ldots\right\}$, and the cumulant generator is

$$\kappa(\theta) = \begin{cases} \log\left(\frac{e^\theta-1}{\theta}\right) & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}. \tag{15}$$

The density of a unifed distribution with canonical parameter θ and dispersion parameter ϕ is

$$f(x;\theta,\phi) = \frac{h(x/\phi,1/\phi)}{\phi}\exp\left(\frac{x\theta-\kappa(\theta)}{\phi}\right), \tag{16}$$

where h and κ are as in (Eq. 14) and (Eq. 15), respectively and $x \in[0,1], \theta \in \mathbb{R}, \phi \in \left\{1, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$. We denote the unifed distribution with canonical parameter θ and dispersion parameter ϕ with $unifed(\theta,\phi)$.

The unifed package does not contain an implementation of (Eq. 16). This is because we did not find a numerically stable way to compute h. To show this, the package includes the function `dirwin.hall` that computes h. Table 1 shows the results we get by calling this function with n set to 50 and varying the values of y. The changes of sign indicate that a float overflow is happening.

Table 1 Float overflow of the Irwin-Hall implementation

Code	Result
dirwin.hall(35,50)	0.0674864
dirwin.hall(36,50)	-13.12745
dirwin.hall(37,50)	45.44388
dirwin.hall(38,50)	-37.44488

The package calls unifed distribution the one-parameter special case of (Eq. 16) where $\phi = 1$, which we denote with *unifed*(θ). This simplifies the density to

$$f(x;\theta) = \begin{cases} \frac{\theta}{e^{\theta}-1} e^{x\theta} & \text{if } \theta \neq 0 \\ 1 & \text{if } \theta = 0 \end{cases} \quad \text{for } x \in (0,1). \tag{17}$$

The functions `dunifed`, `punifed`, `qunifed` and `runifed`, give the density, distribution, quantile and simulation functions, respectively of this simplified version. The mean and variance of each element of the family are given by

$$\mathbb{E}[X] = \dot{\kappa}(\theta) = \begin{cases} \frac{(\theta-1)e^{\theta}+1}{\theta(e^{\theta}-1)} & \text{if } \theta \neq 0 \\ \frac{1}{2} & \text{if } \theta = 0 \end{cases}, \tag{18}$$

$$\mathbb{V}[X] = \dot{\kappa}(\theta) = \begin{cases} \left(\frac{e^{2\theta}-(\theta+2)e^{\theta}+1}{\theta^2(e^{\theta}-1)^2} \right) & \text{if } \theta \neq 0 \\ \frac{1}{12} & \text{if } \theta = 0 \end{cases}, \tag{19}$$

where $\dot{\kappa}$ and $\ddot{\kappa}$ are the first and second derivative of κ, respectively. We have not been able to find an analytical expression for the inverse function $\dot{\kappa}^{-1}$. Thus, it has not been possible either to find analytical expressions for the variance function and unit deviance of the unifed. Nevertheless, the *unifed* package contains the function `unifed.kappa.prime.inverse` that uses the Newthon Raphson method to implement the inverse of $\dot{\kappa}$. This allows us to get a numerical solution for the variance function by using the relation $\mathbf{V}(\mu) = \ddot{\kappa}(\dot{\kappa}^{-1}(\mu))$. This is implemented in the function `unifed.varf`.

Similarly, since the unifed is a regular EDF (see Chapter 2 of (Jørgensen 1997)), we can compute the unit deviance by using the relation

$$d(y,\mu) = 2\left[y\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu))\right]. \tag{20}$$

The function `unifed.unit.deviance` computes the unit deviance using (Eq. 20). As mentioned in "Exponential dispersion families and GLMs" section, the unit deviance can be used to reparametrize the distribution in terms of it's mean and dispersion parameter. We denote with *unifed*$^*(\mu,\phi)$ the unifed distribution with mean μ and dispersion parameter ϕ and when $\phi = 1$, we write simply *unifed*$^*(\mu)$.

Figure 1 shows plots of the unifed distribution for different values of its mean. We can see that except for $\mu = 0.5$, it is always monotone. For $\mu < 0.5$ it is strictly decreasing and the mode is at zero. For $\mu > 0.5$ it is strictly increasing and the mode is at one. The R code used for producing this plot can be found in (Quijano Xacur 2018).

Maximum likelihood estimation

Suppose you have an independent and identically distributed sample $X_1, \ldots, X_n$ coming from a *unifed*(θ) distribution and you want to compute the maximum likelihood estimator (mle) $\hat{\theta}$ of θ. The derivative of the log-likelihood function is given by

$$\begin{aligned} \ell'(\theta|X_1,\ldots,X_n) &= n\frac{(1-\theta)e^{\theta}-1}{\theta(e^{\theta}-1)} + \sum_{i=1}^{n} X_i \\ &= -n\dot{\kappa}(\theta) + \sum_{i=1}^{n} X_i. \end{aligned}$$

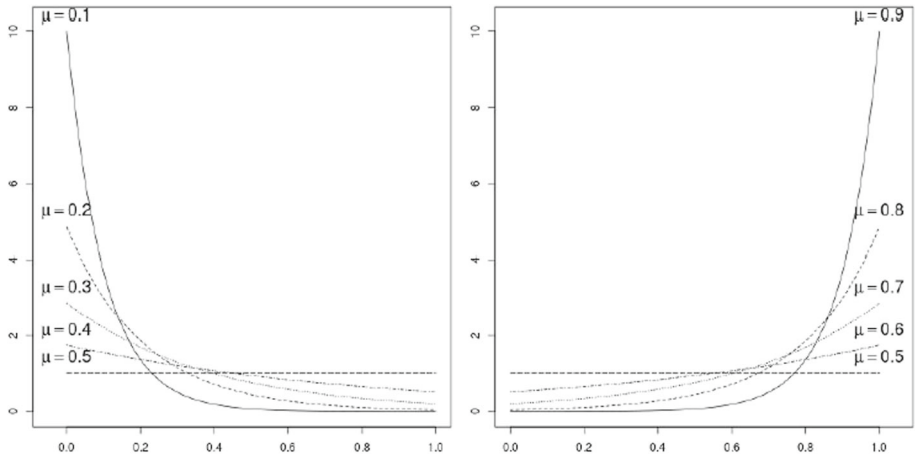

Fig. 1 Density of the unifed for different values of its mean μ

Making the expression above equal to zero and solving for θ, the mle for θ is given by

$$\hat{\theta} = \dot{\kappa}^{-1}\left(\bar{X}\right), \tag{21}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$. The function `unifed.mle` in the unifed R package computes the mle using (Eq. 21). It is possible to use the unifed distribution as the response distribution of a GLM. In this case, ϕ must be fixed to one and the weight of each class is the number of observations in the class. The mle $\hat{\mathcal{B}}$ of the regression coefficients can be found using iterative weighted least squares. In Section 2.5 of (McCullagh and J.A. 1989), they show that this method works for any response distribution whose density can be expressed as (Eq. 8). Thus, the method also works for the unifed. The unifed R package (Quijano Xacur 2019b) provides the function `unifed` that returns a family object than can be used inside the `glm` function.

An illustrative example

In this section we apply a unifed GLM to a publicly available dataset. The data appears in (de Jong and Heller 2008). It is based on 67,856 one–year auto insurance policies from 2004 or 2005. The dataset can be downloaded from the companion site of the book (see (de Jong and Heller 2008)). Table 2 shows the description of the variables as provided at the website.

We are interested in modeling the exposure; which is the proportion of time of the year in which the insurance policy is in-force for a given client. We use `gender`, `agecat`, `area` and `veh_age` as the explanatory variables.

The R code used to obtain the results that follow can be found in (Quijano Xacur 2019a).

The data was aggregated using (Eq. 7) and a unifed GLM was fit to it. Table 3 (exported from R using the package `xtable` (Dahl et al. 2018)) shows the summary provided by the `glm` function of R. We see that all the variables included have at least one significant class.

A χ^2 test for goodness of fit is commonly used for GLMs. The null hypothesis is that the data is distributed according to the fitted GLM. Assuming the null hypothesis for this example implies that the residual deviance reported at the bottom of Table 3 follows a χ^2 distribution with 273 degrees of freedom. The p-value for this example is $\mathbb{P}(\chi^2_{273} \geq 297.86) = 0.14$. Now, the detail with this test is that the χ^2 distribution for the

Table 2 Vehicle insurance variables

Variable name	Description
veh_value	vehicle value, in $10,000s
exposure	0-1
clm	occurrence of claim (0 = no, 1 = yes)
numclaims	number of claims
claimcst0	claim amount (0 if no claim)
veh_body	vehicle body, coded as
	BUS
	CONVT = convertible
	COUPE
	HBACK = hatchback
	HDTOP = hardtop
	MCARA = motorized caravan
	MIBUS = minibus
	PANVN = panel van
	RDSTR = roadster
	SEDAN
	STNWG = station wagon
	TRUCK
	UTE - utility
veh_age	age of vehicle: 1 (youngest), 2, 3, 4
gender	gender of driver: M, F
area	driver's area of residence: A, B, C, D, E, F
agecat	driver's age category: 1 (youngest), 2, 3, 4, 5, 6

Table 3 Summary of Unifed GLM

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.3319	0.0197	-16.84	0.0000 ***
genderM	0.0288	0.0090	3.20	0.0014 **
agecat2	0.0011	0.0184	0.06	0.9518
agecat3	0.0530	0.0178	2.97	0.0029 **
agecat4	0.0583	0.0178	3.28	0.0010 **
agecat5	0.1042	0.0189	5.51	0.0000 ***
agecat6	0.0692	0.0210	3.30	0.0010 ***
areaB	0.0239	0.0135	1.77	0.0761 .
areaC	0.0014	0.0121	0.11	0.9086
areaD	0.0053	0.0157	0.34	0.7337
areaE	0.0120	0.0175	0.68	0.4948
areaF	0.0879	0.0214	4.10	0.0000 ***
veh_age2	0.1708	0.0138	12.40	0.0000 ***
veh_age3	0.1613	0.0133	12.16	0.0000 ***
veh_age4	0.1549	0.0134	11.53	0.0000 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Dispersion parameter for unifed family taken to be 1)				
Null deviance:	585.47 on 287 degrees of freedom			
Residual deviance:	297.86 on 273 degrees of freedom			

residual deviance is asymptotic on the smallest weight of all classes going to infinity (see (Jørgensen 1992, Section 3.6)). The smallest observed weight here is 4 and it corresponds to the class with `gender`=F, `agecat`=6, `area`=F and `veh_age`=1. Therefore the χ^2 test for this example is not reliable.

Figure 2 shows the deviance residuals of this model. It suggests a good fit since they do not show any apparent pattern.

Verifying data aggregation:

We now fit the same model as in the previous section but without aggregating the data. Table 4 shows the summary of the model from R. The code used to generate this table can be found in (Quijano Xacur 2019a).

By comparing Tables 3 and 4 one can see that the estimated coefficients are the same in both cases. Thus, even though the deviance of both models differ, they give the same mle for the coefficients. This shows what we mean with data aggregation.

Comparison between the unifed GLM and the beta regression

The beta regression (Ferrari and Cribari-Neto 2004) is a versatile model for applications with a response variable on the unit interval. Moreover, the well documented R package `betareg` (Cribari-Neto and Zeileis 2010) makes it a practical tool in many applications.

The beta regression

The density of the beta distribution contains a large variety of shapes. In (Ferrari and Cribari-Neto 2004) the beta density is reparameterized as

$$f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1, \tag{22}$$

with $0 < \mu < 1$ and $\phi > 0$, and the distribution is denoted by $\mathcal{B}(\mu, \phi)$. Under this parametrization, if $Y \sim \mathcal{B}(\mu, \phi)$, the mean and variance are

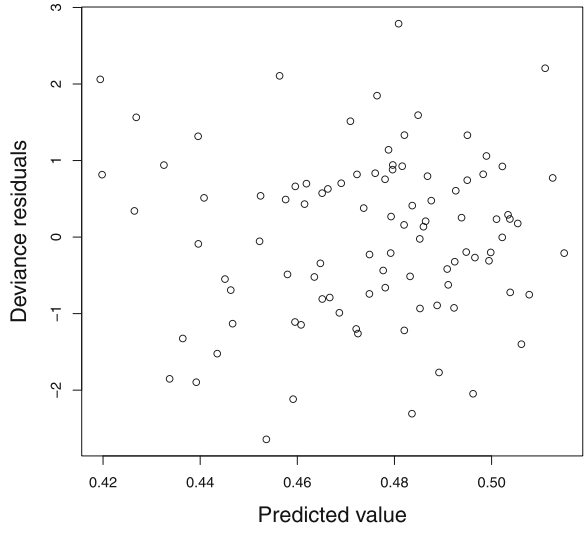

Fig. 2 Residuals of Unifed Regression

Table 4 Summary of Unifed GLM without Data Aggregation

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.3319	0.0197	-16.84	0.0000 ***
genderM	0.0288	0.0090	3.20	0.0014 **
agecat2	0.0011	0.0184	0.06	0.9518
agecat3	0.0530	0.0178	2.97	0.0029 **
agecat4	0.0583	0.0178	3.28	0.0010 **
agecat5	0.1042	0.0189	5.51	0.0000 ***
agecat6	0.0692	0.0210	3.30	0.0010 ***
areaB	0.0239	0.0135	1.77	0.0761 .
areaC	0.0014	0.0121	0.11	0.9086
areaD	0.0053	0.0157	0.34	0.7337
areaE	0.0120	0.0175	0.68	0.4948
areaF	0.0879	0.0214	4.10	0.0000 ***
veh_age2	0.1708	0.0138	12.40	0.0000 ***
veh_age3	0.1613	0.0133	12.16	0.0000 ***
veh_age4	0.1549	0.0134	11.53	0.0000 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Dispersion parameter for unifed family taken to be 1)				
Null deviance:	113445 on 67855 degrees of freedom			
Residual deviance:	113158 on 67841 degrees of freedom			

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}[Y] = \frac{\mu(1-\mu)}{1+\phi}. \tag{23}$$

Here ϕ is called the precision parameter of the distribution. In the beta regression model it is assumed that the response variable is a vector $Y = (Y_1, \ldots, Y_m)$, in which $Y_i \sim \mathcal{B}(\mu_i, \phi)$ for $i = 1, \ldots, m$. The $Y_i's$ are assumed independent to each other. The explanatory variables are incorporated to the model through the relation

$$g(\mu_i) = {x_i}^T \mathcal{B},$$

where $\mathcal{B}$ is a vector of parameters and x_i is a vector of regressors. $g : (0,1) \to \mathbb{R}$ is invertible and is called the link function.

Then (Simas et al. 2010) generalized this model to allow the precision parameter to vary among classes in a similar way to the double generalized linear models (see (Smyth and Verbyla 1999)). More specifically, in this case the response vector $Y = (Y_1, \ldots, Y_m)$ is such that $Y_i \sim \mathcal{B}(\mu_i, \phi_i)$, independently and

$$g_1(\mu_i) = {x_i}^T \mathcal{B},$$
$$g_2(\phi_i) = {z_i}^T \gamma,$$

where $\mathcal{B}$ and γ are regression coefficients.

These regression models offer great flexibility when the response variable lies in the interval $(0,1)$, and both are implemented in the R package `betareg` ((R Core Team 2017), (Cribari-Neto and Zeileis 2010)).

The beta distribution is not an EDF and therefore the beta regression is not a GLM. Nevertheless the parametrization chosen by the authors of the model along with (Eq. 23) give it a similar look and feel.

On the difficulties of data aggregation for the beta regression

Data aggregation gives a practical advantage when working with large datasets. For GLMs this is straightforward due to two properties of $\bar{Y}$ in (Eq. 7):

- $\bar{Y}$ is a sufficient statistic for μ
- The distribution of $\bar{Y}$ belongs to the same family as the Y_i's in (Eq. 7).

We do not know any statistic with these two properties for the beta distribution. For instance, let $Y_1, \ldots, Y_n$ be an i.i.d sample from a $\mathcal{B}(\mu, \phi)$ distribution. The joint likelihood function of this sample is then

$$f(y) = \left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\right)^n \left(\prod_{i=1}^n y_i\right)^{\mu\phi-1} \left(\prod_{i=1}^n (1-y_i)\right)^{(1-\mu)\phi-1},$$

where $y = (y_1, \ldots, y_n)$. This density can be rearranged as follows

$$f(y) = \left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\right)^n \left[\prod_{i=1}^n \frac{(1-y_i)^{\phi-1}}{y_i}\right] \left(\prod_{i=1}^n \frac{y_i}{1-y_i}\right)^{\mu\phi}$$

The factorization theorem (see (Hogg et al. 2019, Chapter 7)), implies that $T = \prod_{i=1}^n \frac{y_i}{1-y_i}$ is sufficient for μ. Now, the distribution of T, which is not beta, would be needed to use T for data aggregation. In other words, a regression model whose response distribution is a family that includes the distribution of T for every n would need to be developed.

Differences between the unifed GLM and the beta regression

The unifed density does not have the variety of shapes that the beta density has. To see this, compare the shapes shown in Figure 1 with the shapes for the beta distribution shown in Figure 3 (Quijano Xacur 2019a). Thus, the beta regression is able to adapt to more shapes than a unifed GLM and even more so if regressors are used for the dispersion parameter.

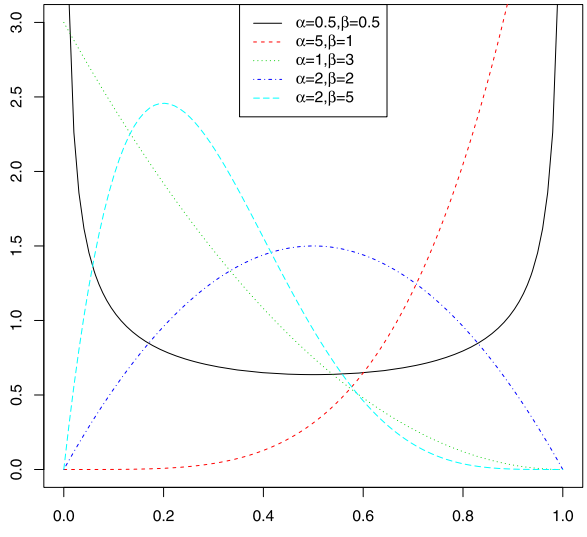

Fig. 3 Beta Densities for Different Values of α and $\mathcal{B}$

In those cases where a beta regression and a unifed GLM give similar good fit, the parsimony principle suggests to pick the unifed GLM, since it has one parameter less; the dispersion parameter is known for the unifed GLM.

From a numerical point of view, the unifed GLM has the advantage that it is possible to use (Eq. 7) for data reduction. This is a practical advantage when dealing with large datasets specially if simulations of the response vector need to be performed.

Conclusion

This paper introduced a new distribution called unifed. It is the Exponential Dispersion Family generated by the uniform distribution. It allows to fit a GLM for responses on the unit interval (0,1). An R package for working with this distribution is provided.

We made a comparison to the beta regression, which is another regression model for responses on the unit interval. It provides more flexible shapes and therefore it can give better fit than a unifed GLM in many situations. In contrast, the unifed GLM is suitable for data aggregation which is a practical advantage when working with large datasets.

An application using publicly available data was presented.

Abbreviations
EDF: Exponential dispersion family; GLM: Generalized linear model; mle: Maximum likelihood estimator

Acknowledgements
Not applicable.

Authors' contributions
All contributions were made by the author of the article, Oscar Alberto Quijano Xacur.

Funding
Not applicable.

Availability of data and material
The data used for the example in this article is publicly available and it can be downloaded from www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/acst_docs/glms_for_insurance_data/data/car.csv.

Competing interests
The author declares that they have no competing interests.

Received: 23 May 2019 Accepted: 24 September 2019
Published online: 05 November 2019

References

Cribari-Neto, F., Zeileis, A.: Beta regression in R. J. Stat. Softw. **34**(2), 1–24 (2010)
Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., Swinton, J.: Xtable: Export Tables to LaTeX or HTML. R package version 1.8-3. (2018). https://CRAN.R-project.org/package=xtable. Accessed Mar 2019
de Jong, P., Heller, G. Z.: Generalized Linear Models for Insurance Data. Cambridge University Press (2008). Companion website: http://www.acst.mq.edu.au/GLMsforInsuranceData. http://dx.doi.org/10.1017/CBO9780511755408
Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**(7), 799–815 (2004). https://doi.org/10.1080/0266476042000214501
Hogg, R. V., McKean, J. W., Craig, A. T.: Introduction to Mathematical Statistics. 8th. Pearson, Boston (2019)
Johnson, N. L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, Vol. 2. Wiley & Sons, New York (1995)
Jørgensen, B.: The Theory of Exponential Dispersion Models and Analysis of Deviance. Instituto de Matemática Pura e Aplicada, (IMPA), Brazil (1992)
Jørgensen, B.: The Theory of Dispersion Models. Chapman & Hall, London (1997)
Smyth, GK, Verbyla, AP: Double generalized linear models: approximate reml and diagnostics. Proceedings of the 14th International Workshop on Statistical Modelling, 66–80 (1999). https://pdfs.semanticscholar.org/3fd5/fb7ee7e6991d0e6e2f50dacc80283a4701b1.pdf
McCullagh, P., J.A., N.: Generalized Linear Models. 2nd. Chapman and Hall, London New York (1989)
Quijano Xacur, O. A.: Beta Density Plot. Code Snippet (2019). https://gitlab.com/oquijano/unifed/snippets/1880287. Accessed Jul 2019
Quijano Xacur, O.A.: Unifed Density Plot. Code Snippet (2018). https://gitlab.com/oquijano/unifed/snippets/1786224. Accessed Jul 2019

Quijano Xacur, O. A.: Vehicle Insurance Example. Code Snippet (2019a). https://gitlab.com/oquijano/unifed/snippets/1786226. Accessed Jul 2019
Quijano Xacur, O.A.: unifed. R package version 1.1.0 (2019b). https://CRAN.R-project.org/package=unifed. Accessed Jul 2019
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). https://www.R-project.org/
Simas, A. B., Barreto-Souza, W., Rocha, A. V.: Improved estimators for a general class of beta regression models. Comput. Stat. Data Anal. **54**(2), 348–366 (2010). https://doi.org/10.1016/j.csda.2009.08.017
Stan Development Team: RStan: the R interface to Stan. R package version 2.18.2 (2018). http://mc-stan.org/

Publisher's Note