Beta regression modeling: recent advances in theory and applications

Silvia L. P. Ferrari

February, 2013

13ª Escola de Modelos de Regressão - Maresias - SP

Introduction

Data measured in a continuous scale and restricted to the unit interval, i.e. $0 < y < 1$:

- percentages,
- proportions,
- fractions,
- rates.

Introduction

Examples

- percentage of time devoted to an activity;
- fraction of income spent on food;
- unemployment rate, poverty rate, etc.;
- math score, reading score, etc.;
- Gini's index;
- fraction of "good cholesterol" (HDL/total cholesterol);
- proportion of sand in the soil;
- fraction of surface covered by vegetation.

More later...

Introduction

Limited range variables usually display:

- heteroskedasticity (the variance is smaller near the extremes);
- asymmetry.

A possible solution for regression analysis is to employ a transformation in the response, and use normality assumption. For example, $\widetilde{y} = \log[y/(1-y)]$.

Drawbacks:

- difficult to correct for both heteroskedasticity and asymmetry;
- the model parameters cannot be easily interpreted in terms of the original response.

Introduction

It is more appropriate to use a regression model that assumes that the response variable follows a continuous distribution with support in $(0, 1)$. Example: Beta regression.

Beta regression models employ a parameterization of the beta distribution in terms of its mean and a precision (or dispersion) parameter.

Simple beta regression models are similar to generalized linear models.

Some early references: Paolino (2001); Kieschnick & McCullough (2003); Ferrari & Cribari–Neto (2004); Cepeda & Gamerman (2005).

Beta regression models

Beta density:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1,$$

where $0 < \mu < 1$ and $\phi > 0$. Note that

$$\mathrm{E}(y) = \mu$$

and

$$\mathrm{var}(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Hence, ϕ can be regarded as a precision parameter.

This is not the usual parameterization of the beta law, but is convenient for modeling purposes.

Beta regression models

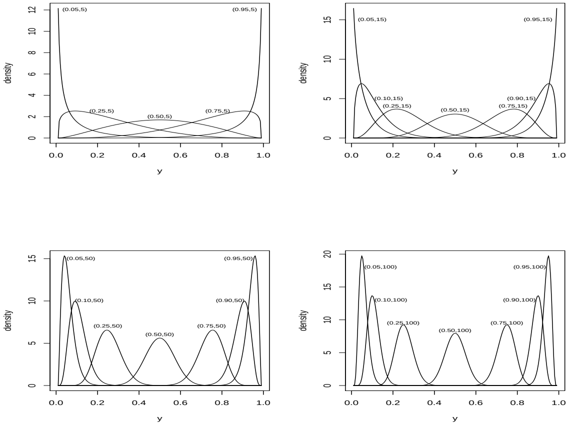

Figure 1. Beta densities for different combinations of (μ, ϕ).

A simple beta regression model

Ref.: Ferrari & Cribari–Neto (2004)

- $y_1, \ldots, y_n$: independent r.v.;
- y_t, $t = 1, \ldots, n$, follow a **beta distribution** with **mean** μ_t and unknown **precision** ϕ, i.e. $y_t \sim \mathrm{Beta}(\mu_t, \phi)$;
- $g(\cdot)$: strictly monotone and twice differentiable **link function** that maps $(0, 1)$ to $\mathbb{R}$.
- $g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t$;
- $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}$ is a vector of unknown regression parameters;
- $x_{t1}, \ldots, x_{tk}$ are observations on k covariates ($k < n$).

A simple beta regression model

Some possible link functions:

1. Logit: $g(\mu) = \log[\mu/(1-\mu)]$;
2. Probit: $g(\mu) = \Phi^{-1}(\mu)$;
3. Complimentary log–log: $g(\mu) = \log[-\log(1-\mu)]$;
4. Log-log: $g(\mu) = -\log[-\log(\mu)]$;
5. Cauchit: $g(\mu) = \tan[\pi(\mu - 0.5)]$.

- The link functions above are the inverse cumulative distribution functions (quantile functions) of well-known distributions (1. logistic, 2. standard normal, 3. minimum extreme value, 4. maximum extreme value, 5. Cauchy).
- For a discussion on link functions, see Ramalho, Ramalho & Murteira (2010).

A simple beta regression model

Log-likelihood:

$$\ell(\beta, \phi) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi),$$

where

$$\begin{aligned} \ell_t(\mu_t, \phi) &= \log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1-\mu_t)\phi) \\ &\quad + (\mu_t \phi - 1) \log y_t + \{(1-\mu_t)\phi - 1\} \log(1 - y_t). \end{aligned}$$

Let

$$y_t^* = \log \frac{y_t}{1 - y_t}$$

and

$$\mu_t^* = \mathrm{E}(y_t^*) = \psi(\mu_t \phi) - \psi((1-\mu_t)\phi).$$

A simple beta regression model

Score function for β:

$$U_\beta(\beta, \phi) = \phi X^\top T(y^* - \mu^*),$$

where X is an $n \times k$ matrix whose t-th row is $x_t^\top$,

$$T = \mathrm{diag}\{1/g'(\mu_1)\ldots, 1/g'(\mu_n)\},$$

$y^* = (y_1^*, \ldots, y_n^*)^\top$ and $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^\top$.

Score function for ϕ:

$$\begin{aligned} U_\phi(\beta, \phi) &= \sum_{t=1}^{n} \{\mu_t(y_t^* - \mu_t^*) + \log(1 - y_t) \\ &- \psi((1 - \mu_t)\phi) + \psi(\phi)\}. \end{aligned}$$

A simple beta regression model

Fisher information:

$$K = K(\beta,\phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix},$$

where $K_{\beta\beta} = \phi X^\top W X$, $K_{\beta\phi} = K_{\phi\beta}^\top = X^\top Tc$, $K_{\phi\phi} = \mathrm{tr}(D)$, $W = \mathrm{diag}\{w_1,\ldots,w_n\}$, $c = (c_1,\ldots,c_n)^\top$ and $D = \mathrm{diag}\{d_1,\ldots,d_n\}$, with

$$w_t = \phi\left\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\right\}\frac{1}{\{g'(\mu_t)\}^2},$$

$$c_t = \phi\left\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\right\},$$

$$d_t = \psi'(\mu_t\phi)\mu_t^2 + \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi).$$

β **and** ϕ **are not orthogonal**.

A simple beta regression model

In large samples,

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1} \right),$$

approximately, where $\hat{\beta}$ and $\hat{\phi}$ are the MLEs of β and ϕ, and

$$K^{-1} = K^{-1}(\beta, \phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix},$$

where

$$K^{\beta\beta} = \frac{1}{\phi}(X^\top WX)^{-1}\left\{ I_k + \frac{X^\top Tcc^\top T^\top X(X^\top WX)^{-1}}{\gamma\phi} \right\},$$

with $\gamma = \mathrm{tr}(D) - \phi^{-1}c^\top T^\top X(X^\top WX)^{-1}X^\top Tc$,

$$K^{\beta\phi} = (K^{\phi\beta})^\top = -\frac{1}{\gamma\phi}(X^\top WX)^{-1}X^\top Tc, \quad K^{\phi\phi} = \gamma^{-1}.$$

A simple beta regression model

- **MLEs**: obtained numerically by maximizing the log-likelihood function.
- Ferrari & Cribari–Neto (2004) suggest reasonable initial estimates for β and ϕ.
- Also in Ferrari & Cribari–Neto (2004): some simple diagnostic tools and applications.
- `R` implementation of beta regression inference and diagnostics: `betareg` package (Cribari–Neto & Zeiles, 2010).
- A Bayesian approach to beta regression: Branscum, Johnson & Thurmond (2007).

More general beta regression models

- Varying dispersion beta regression models: the mean and the precision parameters are modeled through linear regression structures. Smithson & Verkuilen (2006).
- A general class of beta regression models: the mean and the precision parameters are modeled through linear or nonlinear regression structures. Simas, Barreto-Souza & Rocha (2010).
- Inflated beta regression models: allow zero and/or one occurrences by incorporating degenerate distributions to model the extreme values. Cook, Kieschnick, McCullough (2008), Ospina & Ferrari (2010, 2012a), Calabrese (2012).
- Truncated inflated beta regression models: allow truncation in a subset $[c, 1]$ of the unit interval, and mass points at c, zero and one. Pereira, Botter & Sandoval (2011, 2013).

More general beta regression models

- Semi-parametric beta regression: Branscum, Jonhson & Thurmond (2007), Weihua et al (2012).
- Time series: Rydlewski (2007), Rocha & Cribari–Neto (2009), Billio & Casarin (2011), Casarin, Dalla Valle, Leisen (2012); da-Silva, Migon & Correia (2011), da-Silva & Migon (2012), Guolo & Varin (2012).
- Multivariate beta regression: Souza & Moura (2012a, 2012b)
- Mixed beta regression: Zimprich (2010), Verkuilen & Smithson (2012), Figueroa–Zúñiga, Arellano–Valle & Ferrari (2013), Bonat, Ribeiro Jr & Zeviani (2013).
- Errors-in-variables beta regression models: Carrasco, Ferrari, Arellano–Valle (2012) (more later).
- Beta rectangular regression models: Bayes, Bazán & García (2012).

Special topics in beta regression

- Diagnostics:
  - Espinheira, Ferrari, Cribari–Neto (2008a, 2008b) and Chien (2011, 2012) [beta regression with constant precision]
  - Ferrari, Espinheira & Cribari–Neto (2011), Rocha & Simas (2011) [varying dispersion/nonlinear beta regression models]
  - Anholeto, Sandoval & Botter (2012) [beta regression with constant precision; adjusted residuals]
- Specification tests: Ramalho, Ramalho & Murteira (2010) [with review on models for fractional data], Pereira & Cribari–Neto (2013).
- Robust inference in varying dispersion beta regression: Cribari–Neto & Souza (2012).
- Optimal designs: Wu, Fedorov & Propert (2005).

Special topics in beta regression

- Consistency and asymptotic normality of MLEs: Rydlewski & Mielczarek (2012).
- Bias correction of MLEs:
  - Ospina, Cribari–Neto, Vasconcellos (2006) and Kosmidis & Firth (2010) [beta regression with constant precision]
  - Simas, Barreto-Souza & Rocha (2010) [general beta regression]
  - Ospina & Ferrari (2012b) [inflated beta regression]
- Size-corrected tests:
  - Ferrari & Pinheiro (2011) [general beta regression; Skovgaard's adjustment]
  - Bayer & Cribari–Neto (2012) [simple beta regression; Bartlett correction]
  - Cribari–Neto & Queiroz (2012) [varying dispersion beta regression; Skovgaard, bootstrap, comparison among various tests]
  - Pereira & Cribari–Neto (2012) [inflated beta regression; Skovgaard]

Errors-in-variables beta regression

Ref.: Carrasco, Ferrari, Arellano–Valle (2012)

- $y_t \sim \mathrm{Beta}(\mu_t, \phi_t)$, for $t = 1, \ldots, n$;
- mean and precision submodels:

$$g(\mu_t) = \mathbf{z}_t^\top \boldsymbol{\alpha} + \mathbf{x}_t^\top \boldsymbol{\beta}, \quad (1)$$
$$h(\phi_t) = \mathbf{v}_t^\top \boldsymbol{\gamma} + \mathbf{m}_t^\top \boldsymbol{\lambda}; \quad (2)$$

- $\boldsymbol{\alpha} \in \mathbb{R}^{p_\alpha}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_\beta}$, $\boldsymbol{\gamma} \in \mathbb{R}^{p_\gamma}$, $\boldsymbol{\lambda} \in \mathbb{R}^{p_\lambda}$ are column vectors of unknown parameters;
- $\mathbf{x}_t = (x_{t1}, \cdots, x_{tp_\beta})^\top$ and $\mathbf{m}_t = (m_{t1}, \cdots, m_{tp_\lambda})^\top$ are unobservable (observed with error) covariates;
- the vectors of covariates measured without error, $\mathbf{z}_t$ and $\mathbf{v}_t$, may contain variables in common, and likewise, $\mathbf{x}_t$ and $\mathbf{m}_t$.
- given the covariates, $y_1, \ldots, y_n$ are assumed to be independent.

Errors-in-variables beta regression

- $\mathbf{s}_t$: vector containing all the unobservable covariates;
- $\mathbf{w}_t$ is observed in place of $\mathbf{s}_t$;
- it is assumed that

$$\mathbf{w}_t = \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 \circ \mathbf{s}_t + \mathbf{e}_t, \tag{3}$$

  where $\mathbf{e}_t$ is a vector of random errors, $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$ are (possibly unknown) parameter vectors and $\circ$ is the element-wise product;
- $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$: additive and multiplicative biases of the measurement error mechanism, respectively;
- classical additive model: $\mathbf{w}_t = \mathbf{s}_t + \mathbf{e}_t$;
- we follow the structural approach; the unobservable covariates are regarded as random variables;
- we assume that $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are iid;
- it is assumed that they are independent of the measurement errors $\mathbf{e}_1, \ldots, \mathbf{e}_n$;
- the normality assumption for the joint distribution of $\mathbf{s}_t$ and $\mathbf{e}_t$ is assumed;
- parameters of the joint distribution of $\mathbf{w}_t$ and $\mathbf{s}_t$: δ.

Errors-in-variables beta regression

- $(y_1, \mathbf{w}_1), \ldots, (y_n, \mathbf{w}_n)$: observable variables.
- We omit the observable vectors $\mathbf{z}_t$ and $\mathbf{v}_t$ in the notation as they are non-random and known.
- The joint density of $(y_t, \mathbf{w}_t)$ is obtained by integrating the joint density of the complete data $(y_t, \mathbf{w}_t, \mathbf{s}_t)$,

$$f(y_t, \mathbf{w}_t, \mathbf{s}_t; \theta, \delta) = f(y_t|\mathbf{w_t}, \mathbf{s}_t; \theta) f(\mathbf{s}_t, \mathbf{w}_t; \delta),$$

  with respect to $\mathbf{s}_t$.
- $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ represents the parameter of interest, and $\boldsymbol{\delta}$ is the nuisance parameter.
- The joint density associated to the measurement error model, $f(\mathbf{w}_t, \mathbf{s}_t; \delta)$, can be written as $f(\mathbf{w}_t, \mathbf{s}_t; \delta) = f(\mathbf{w}_t|\mathbf{s}_t; \delta) f(\mathbf{s}_t|\delta)$ as well as $f(\mathbf{w}_t, \mathbf{s}_t; \delta) = f(\mathbf{s}_t|\mathbf{w}_t; \delta) f(\mathbf{w}_t|\delta)$.

Errors-in-variables beta regression

- We assume that, given the true (unobservable) covariates $\mathbf{s}_t$, the response variable y_t does not depend on the surrogate covariates $\mathbf{w}_t$; i.e. $f(y_t|\mathbf{w}_t, \mathbf{s}_t; \theta) = f(y_t|\mathbf{s}_t; \theta)$.
- The density function of $(y_t, \mathbf{w}_t)$ is given by

$$
\begin{aligned}
f(y_t, \mathbf{w}_t; \theta, \delta) &= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(y_t, \mathbf{w}_t, \mathbf{s}_t; \theta, \delta) d\mathbf{s}_t, \\
&= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(y_t|\mathbf{s}_t; \theta) f(\mathbf{w}_t, \mathbf{s}_t; \delta) d\mathbf{s}_t.
\end{aligned}
$$

- Log-likelihood function:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \sum_{t=1}^{n} \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_t|\mathbf{s}_t; \boldsymbol{\theta}) f(\mathbf{s}_t|\mathbf{w}_t; \boldsymbol{\delta}) f(\mathbf{w}_t; \boldsymbol{\delta}) d\mathbf{s}_t, \\
&= \sum_{t=1}^{n} \log f(\mathbf{w}_t; \boldsymbol{\delta}) + \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_t|\mathbf{s}_t; \boldsymbol{\theta}) f(\mathbf{s}_t|\mathbf{w}_t; \boldsymbol{\delta}) d\mathbf{s}_t.
\end{aligned}
$$

- The likelihood function involves analytically intractable integrals.
- Approximate inference methods needed.

Errors-in-variables beta regression

In order to facilitate the description of the estimation methods, we consider the following model:

$$y_t|x_t, x_t \sim \mathrm{Beta}(\mu_t, \phi_t),$$
$$g(\mu_t) = \mathbf{z}_t^\top \boldsymbol{\alpha} + x_t\beta, \quad h(\phi_t) = \mathbf{v}_t^\top \gamma + x_t\lambda,$$
$$w_t = \tau_0 + \tau_1 x_t + e_t, \;\; x_t \overset{\text{ind}}{\sim} N(\mu_x, \sigma_x^2), \;\; e_t \overset{\text{ind}}{\sim} N(0, \sigma_e^2),$$

with x_t and $e_{t'}$, for $t, t' = 1, \ldots, n$, being independent. The unknown parameter vectors $\boldsymbol{\alpha}$ and γ were defined above, and $\beta \in \mathbb{R}$, $\lambda \in \mathbb{R}$, $\mu_x \in \mathbb{R}$ and $\sigma_x^2 > 0$ are unknown parameters.

Errors-in-variables beta regression

We have

$$w_t \overset{\text{ind}}{\sim} \mathrm{N}(\tau_0 + \tau_1 \mu_x, \tau_1^2 \sigma_x^2 + \sigma_e^2), \quad x_t | w_t \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{x_t|w_t}, \sigma^2_{x_t|w_t}),$$

where

$$\mu_{x_t|w_t} = \mu_x + k_x[w_t - (\tau_0 + \tau_1 \mu_x)], \quad \sigma^2_{x_t|w_t} = \sigma_e^2 k_x / \tau_1,$$

with $k_x = \tau_1 \sigma_x^2 / (\tau_1^2 \sigma_x^2 + \sigma_e^2)$ being known as the reliability ratio.

To avoid non-identifiability of parameters we assume that $(\tau_0, \tau_1, \sigma_e^2)$ or (τ_0, τ_1, k_x) are either known parameters or they are estimated from supplementary information, typically replicate measurements or partial observation of the error-free covariate.

In any case, either of these vectors are regarded as known quantities in the inferential procedure. Hence, the nuisance parameter vector is $\boldsymbol{\delta} = (\mu_x, \sigma_x^2)^\top$.

Errors-in-variables beta regression

Log-likelihood function:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{t=1}^{n} \ell_{1t}(\boldsymbol{\delta}) + \sum_{t=1}^{n} \ell_{2t}(\boldsymbol{\theta}, \boldsymbol{\delta}),$$

where

$$\ell_{1t}(\boldsymbol{\delta}) = -\frac{1}{2}\log[2\pi(\tau_1^2\sigma_x^2 + \sigma_e^2)] - \frac{[w_t - (\tau_0 + \tau_1\mu_x)]^2}{2(\tau_1^2\sigma_x^2 + \sigma_e^2)},$$

$$\ell_{2t}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \log \int_{-\infty}^{\infty} f(y_t|x_t; \boldsymbol{\theta}) \frac{1}{\sqrt{2\pi\sigma^2_{x_t|w_t}}} \exp\left[-\frac{(x_t - \mu_{x_t|w_t})^2}{2\sigma^2_{x_t|w_t}}\right] dx_t.$$

Errors-in-variables beta regression

Carrasco, Ferrari & Arellano–Valle present three estimation methods:

- Approximate maximum likelihood
  - $\ell_{2t}(\boldsymbol{\theta}, \boldsymbol{\delta})$ is approximated using the Gauss-Hermite quadrature, resulting in an approximate log-likelihood function, $\ell_a(\boldsymbol{\theta}, \boldsymbol{\delta})$;
  - the estimator of $(\boldsymbol{\theta}^\top, \boldsymbol{\delta}^\top)^\top$, $(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\delta}}^\top)^\top$ say, is obtained by solving the system of equations $\partial\ell_a(\boldsymbol{\theta}, \boldsymbol{\delta})/\partial\boldsymbol{\theta} = 0$;
  - for n and Q (the number of quadrature points) sufficiently large, $(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\delta}}^\top)^\top$ is approximately normally distributed with mean $(\boldsymbol{\theta}^\top, \boldsymbol{\delta}^\top)^\top$ and covariance matrix $\mathbf{J}_a^{-1}(\boldsymbol{\theta}, \boldsymbol{\delta})$, where

$$\mathbf{J}_a(\boldsymbol{\theta}, \boldsymbol{\delta}) = -\frac{\partial^2\ell_a(\boldsymbol{\theta}, \boldsymbol{\delta})}{\partial(\boldsymbol{\theta}^\top, \boldsymbol{\delta}^\top)^\top\partial(\boldsymbol{\theta}^\top, \boldsymbol{\delta}^\top)}$$

    Guolo (2011);
  - for computational implementation, the derivatives of $\ell_a(\boldsymbol{\theta}, \boldsymbol{\delta})$ with respect to the parameters can be analytically obtained or numerical derivatives can be used.

Errors-in-variables beta regression

- Approximate maximum pseudo-likelihood
  - The nuisance parameter vector $\boldsymbol{\delta}$ is estimated by maximizing the reduced log-likelihood function

$$\ell_r(\boldsymbol{\delta}) = \sum_{t=1}^{n} \ell_{1t}(\boldsymbol{\delta}).$$

  - The estimate of $\boldsymbol{\delta}$, $\widehat{\boldsymbol{\delta}}$, is inserted in the original log-likelihood function, which results in the pseudo-log-likelihood function

$$\ell_p(\boldsymbol{\theta}; \widehat{\boldsymbol{\delta}}) = \sum_{t=1}^{n} \ell_{1t}(\widehat{\boldsymbol{\delta}}) + \sum_{t=1}^{n} \ell_{2t}(\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}).$$

  - The second term in $\ell_p(\boldsymbol{\theta}; \widehat{\boldsymbol{\delta}})$ is analytically intractable. Unlike the integral in $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$, the integral in $\ell_p(\boldsymbol{\theta}; \widehat{\boldsymbol{\delta}})$ depends on the parameter of interest only.
  - $\ell_{2t}(\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}})$ is approximated using the Gauss-Hermite quadrature.
  - Carrasco, Ferrari & Arellano–Valle (2012) present the limiting distribution of $\widehat{\boldsymbol{\theta}}$, the resulting estimator of $\boldsymbol{\theta}$.

Errors-in-variables beta regression

- Regression calibration estimation
  - Idea: replace the unobservable variable, x_t, by an estimate of $\mathrm{E}(x_t|w_t)$ in the likelihood function.
  - $\mathrm{E}(x_t|w_t) = \mu_{x_t|w_t}$ (calibration function).
  - $\overline{w} = \sum_{t=1}^{n} w_t/n$ and $s_w^2 = \sum_{t=1}^{n}(w_t - \overline{w})^2/(n-1)$ are optimal estimates of $\tau_0 + \tau_1\mu_x$ and $\tau_1^2\sigma_x^2 + \sigma_e^2$, respectively.
  - These estimates can be used to estimate the calibration function.
  - By inserting the estimated calibration function in the conditional density function of y_t given x_t, we obtain a modified log-likelihood function, $\ell_{rc}(\boldsymbol{\theta})$, which equals the log-likelihood function for a beta regression model without errors in covariates and with x_t replaced by $\widetilde{x}_t$, the estimated calibration function.
  - The regression calibration estimate of $\boldsymbol{\theta}$ is obtained from the system of equations $\partial\ell_{rc}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = 0$, which requires a numerical algorithm; e.g. `betareg` package.
  - Numerical evidence indicates that the regression calibration estimator is not consistent.

Errors-in-variables beta regression

Simulation

- Mean submodel: $\log(\mu_t/(1-\mu_t)) = \alpha + \beta x_t$.
- Precision submodels: $\log(\phi_t) = \gamma$ (constant precision model) and $\log(\phi_t) = \gamma + \lambda x_t$ (varying precision model).
- Parameter values: $\alpha =$2.0, $\beta = -0.6$, $\lambda = 0.5$, $\mu_x = 2.5$, $\sigma_x^2 = 2.7$, and $\gamma = 2.5$ (constant precision model) and $\gamma = 4$ (varying precision model).
- Parameters of the measurement error mechanism (known): $\tau_0 = 0$, $\tau_1 = 1$, and values for σ_e^2: 0.05 and 0.50 ($k_x = 0.98$, and 0.84).
- Settings:
  1. we ignored the measurement error in x_t – naïve method – (ℓ_{naive});
  2. we recognized that x_t is measured with error – approximate maximum likelihood (ℓ_a), approximate maximum pseudo-likelihood (ℓ_p), and regression calibration (ℓ_{rc}).
- Number of quadrature points: $Q = 50$.

Errors-in-variables beta regression

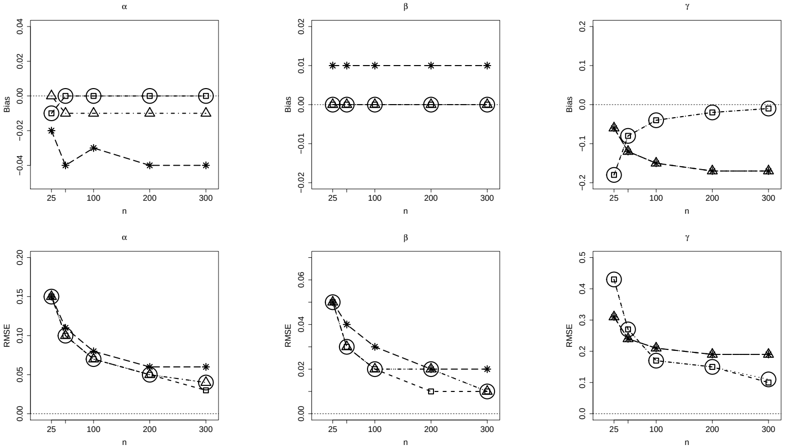

Figure : Bias and RMSE for the estimators of α, β and γ for $k_x = 0.98$, constant precision model; ℓ_a (square), ℓ_p (circle), ℓ_{rc} (triangle) and ℓ_{naive}(star).

Errors-in-variables beta regression

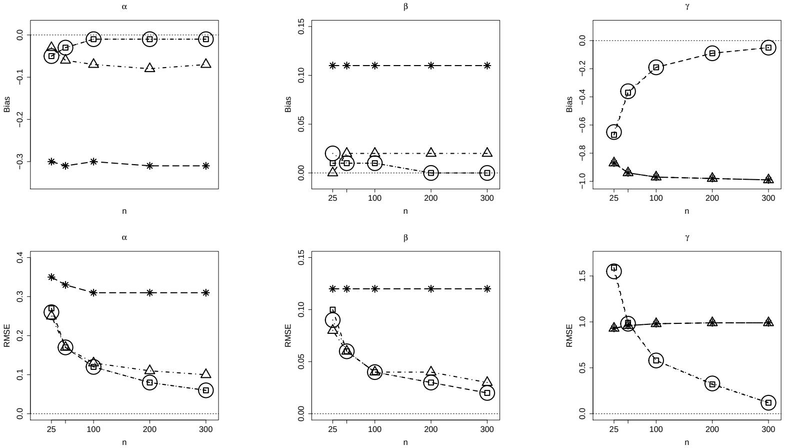

Figure : Bias and RMSE for the estimators of α, β and γ for $k_x = 0.84$, constant precision model; ℓ_a (square), ℓ_p (circle), ℓ_{rc} (triangle) and ℓ_{naive}(star).

Errors-in-variables beta regression

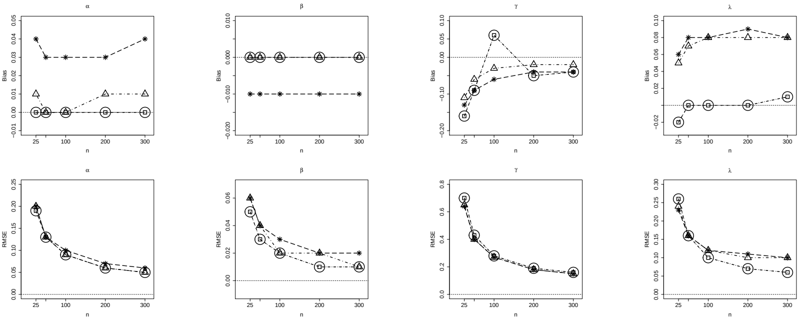

Figure : Bias and RMSE for the estimators of α, β, γ and λ for $k_x = 0.98$, varying precision model; ℓ_a (square), ℓ_p (circle), ℓ_{rc} (triangle) and ℓ_{naive}(star).

Errors-in-variables beta regression

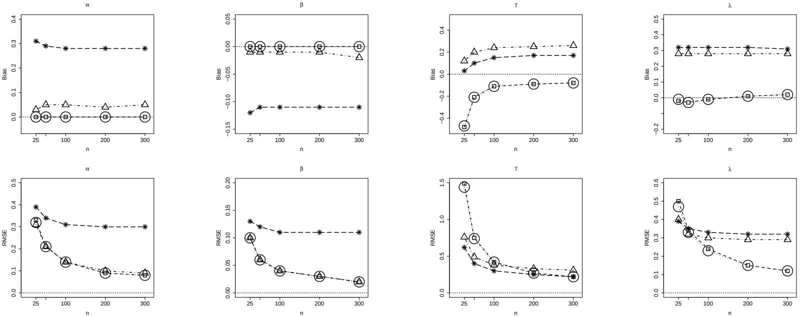

Figure : Bias and RMSE for the estimators of α, β, γ and λ for $k_x = 0.84$, varying precision model; ℓ_a (square), ℓ_p (circle), ℓ_{rc} (triangle) and ℓ_{naive}(star).

Errors-in-variables beta regression

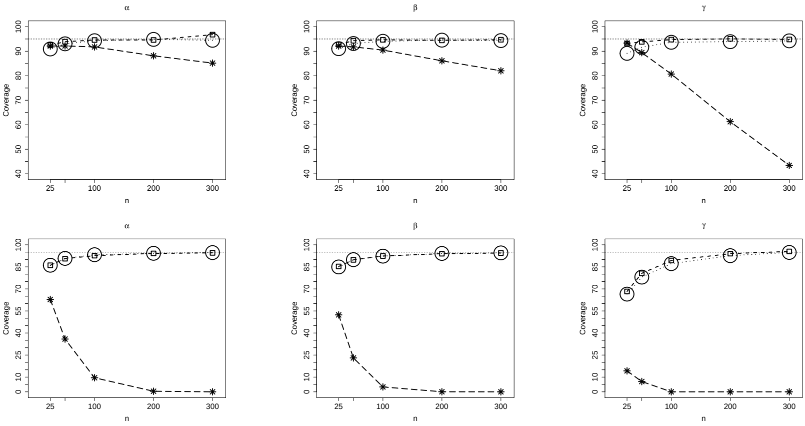

Figure : Coverage of 95% confidence intervals of α, β and γ for: $k_x = 0.98$, constant precision model (first row), and $k_x = 0.84$, constant precision model (second row); ℓ_a (square), ℓ_p (circle), and ℓ_{naive}(star).

Errors-in-variables beta regression

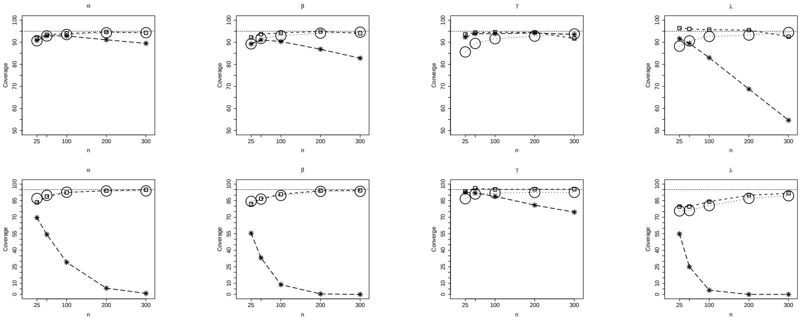

Figure : Coverage of 95% confidence intervals of α, β, γ and λ for: $k_x = 0.98$, varying precision model (first row), and $k_x = 0.84$, varying precision model (second row); ℓ_a (square), ℓ_p (circle), and ℓ_{naive}(star).

Errors-in-variables beta regression

Simulation: Bias and RMSE

- The naïve estimator is clearly not consistent.
- The approximate maximum likelihood and maximum pseudo-likelihood estimators perform similarly.
- Their performance is clearly better than that of the regression calibration and naïve estimators.
- Under constant precision the regression calibration estimator is as biased as the naïve estimator for estimating the precision parameter. For estimating β, the coefficient associated to the covariate measured with error, it performs well if the measurement error variance is small.
- The regression calibration, approximate maximum likelihood and maximum pseudo-likelihood estimators are virtually unbiased for estimating β when $k_x = 0.98$. Their mean-square errors converge to zero as n grows.
- There is evidence that the regression calibration estimator is not consistent.
- Under the varying precision model similar conclusions are reached.

Errors-in-variables beta regression

Simulation: Confidence intervals

- For all the cases, the estimated true coverages of the confidence intervals based on the naïve estimator decrease as n grows. It cannot be recommended.
- The confidence intervals constructed from the approximate maximum likelihood and maximum pseudo-likelihood estimators present true coverage close to 95%, more so if the sample size is large.
- Under the varying precision model, we arrive at similar conclusions.

Errors-in-variables beta regression

Simulation: Conclusion

- Ignoring the measurement error produces misleading inference.
- Inference based on the approximate likelihood and the approximate pseudo-likelihood methods present good performance for the estimation of all the parameters.
- Since the pseudo-likelihood approach is computationally less demanding than the approximate maximum likelihood approach, we recommend the approximate maximum pseudo-likelihood estimation for practical applications.

Also in Carrasco, Ferrari & Arellano–Valle (2012): residual analysis, application.

Applications of beta regression models

- Applications of beta regression are found in various fields.
- I found approximately 100 papers.
- Beta regression is useful and software is available. E.g.
  - R:
    - `betareg` (Cribari–Neto & Zeiles, 2010; Grün, Kosmidis & Zeiles, 2012),
    - `gamlss` (Stasinopoulos & Rigby, 2007)
  - SAS PROC `NLMIXED`: Macro `Beta_Regression` (Swearingen et al., 2011, 2012)
  - examples using R, SPLUS, SAS and SPSS: http://psychology3.anu.edu.au/people/smithson/details/betareg/
- More on computational implementation: Raydonal Ospina later.

Some examples follow.

Applications of beta regression models

Medicine

- proportion of baseline (no glasses) UVB exposure that a person receives if he wears glasses (Egleston et al, 2006)
- measure of lens opacity (cataract) (Chylack Jr et al, 2009)
- proportion of assigned treatment actually taken (Ma, Roy, Marcus, 2010)
- proportion of myocardial necrosis area in patients with acute myocardial infarction (Pinto et al, 2011)
- quality of life measured on a scale of 0-1 of HIV/AIDS patients (Hubben et al. 2008)
- health-related quality of life in stroke patients measured by the Stroke Impact Scale (SIS) (Hunger, Döring, Holle 2012)
- percent mammographic density (high MD is a marker of breast cancer) (Peplonska, 2012)

Applications of beta regression models

Veterinary medicine (genetics)

- genetic difference between two foot-and-mouth virus strains measured as the proportion of nucleotides that differ for a defined portion of the genome (Branscum, Jonhson & Thurmond 2007).

Pharmacology

- score of cognitive impairment in Alzheimer's patients (Rogers et al, 2012) – meta-analysis

Odontology

- percentage of clinical attachment loss (CAL) $\geq 3.5\mathrm{mm}$ and $\geq 7.0\mathrm{mm}$ (CAL measured at six sites per tooth) (Abdo et al, 2012)

Hydrobiology

- fraction of organic matter of the total suspended particulate matter in a sampling zone of a river (Wallis, 2009)

Applications of beta regression models

Aquaculture nutrition

- protein and lipid egg content ($\mathrm{g\,kg}^{-1}$) of female channel catfish (Quintero et al, 2011)

Forest Science

- percent canopy cover (Korhonen et al, 2007)
- percent shrub cover (Ekleston et al, 2011)

Education

- score of educational performance (Carmichael, 2006)
- score of reading accuracy (Smithson & Verkuilen, 2006)

Political Science

- percentage of individuals who feel that race is the most important problem facing America (Gillion, 2008)

Applications of beta regression models

Economics

- percentage of females in municipal councils and executive committees (De Paola, Scoppa, Lombardo, 2010)
- proportion of total annual Asian Development Bank lending committed to a particular country for environmentally risky (non-risky) projects (Buntaine, 2011)
- central-bank independence measured in terms of an index bounded between 0 and 1 (Berggren, Daunfeldt & Hellströn, 2012)
- Artist Price Heterogeneity (APH) measured by Gini's index calculated on artist price distribution (Castellani, Pattitoni & Scorcu 2012)

Applications of beta regression models

Credit risk

- Loss given default (LGD) (Huang & Oosterlee, 2011)

Waste management

- municipal waste separation rates in Spanish cities (Ibáñez, Prades & Simó, 2011; Gallardo et al, 2012)

Social Science

- subjective survival probability (SSP) derived from the question "What are the chances that you will live to be age T or more?" (Balia, 2011)

Conclusion

- Beta regression is useful for practical applications.
- Growing literature on beta regression over the last few years.
- Computational implementation is available.
- There is room for new research.

References

Theory

1. Anholeto, T., Sandoval, D.A & Botter, D.A. (2012). Adjusted Pearson residuals in beta regression models. *Journal of Statistical Computation and Simulation*. In press. DOI: 10.1080/00949655.2012.736993
2. Bayer, F.M. & Cribari, F. (2012). Bartlett corrections in beta regression models. *Journal of Statistical Planning and Inference*, **143**, 531–547.
3. Bayes, C., Bazán, J.L. & García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 771–796.
4. Billio, M. & Casarin, R. (2011). Beta autoregressive transition Markov-switching models for business cycle analysis. *Studies in Nonlinear Dynamics & Econometrics*, **15**, Article 2.
   Bonat, W.H., Ribeiro Jr, P.J. & Zeviani, W.M. (2013). Likelihood analysis for a class of beta mixed models. Not published.
5. Branscum, A.J., Johnson, W.O. & Thurmond, M.C. (2007). Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, **49**, 287–301.
6. Calabrese, R. (2012). Regression model for proportions with probability masses at zero and one. *Working Paper*. Available at http://www.ucd.ie/geary/static/publications/workingpapers/gearywp201209.pdf.
7. Carrasco, J.M.F., Ferrari, S.L.P. & Arellano–Valle, R.B. (2012). Errors-in-variables beta regression models. Available at arXiv:1212.0870.
8. Casarin, R., Dalla Valle, L. & Leisen, F. (2012). Bayesian model selection for beta autoregressive processes. *Bayesian Analysis*, **7**, 1–26.
9. Cepeda Cuervo, E. & Gamerman, D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística*, **57**, 93–105.
10. Chien, L.-C. (2011). Diagnostic plots in beta-regression models. *Journal of Applied Statistics*, **38**, 1607–1622.
11. Chien, L.-C. (2012). Multiple deletion diagnostics in beta regression models. *Computational Statistics*. DOI: 10.1007/s00180-012-0370-9.
12. Cook, D.O., Kieschnick, R., McCullough, B.D. (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, **15**, 860–867.
13. Cribari–Neto, F. & Souza, T.C. (2012). Testing inference in variable dispersion beta regressions. *Journal of Statistical Computation and Simulation*, **82**, 1827–1843.
14. Cribari–Neto, F. & Queiroz, M.P.F. (2012). On testing inference in beta regressions. *Journal of Statistical Computation and Simulation*. DOI:10.1080/00949655.2012.700456.
15. da-Silva, C.Q., Migon, H.S. & Correia, L.T. (2011). Dynamic Bayesian beta models. *Computational Statistics and Data Analysis*, **55**, 2074–2089.

16. da-Silva, C.Q. & Migon, H.S. (2012). Hierarchical dynamic beta model. Technical report n. 253. Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro.

17. Espinheira, P.L., Ferrari, S.L.P. & Cribari–Neto, F. (2008a). Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, **52**, 4417–4431.

18. Espinheira, P.L., Ferrari, S.L.P. & Cribari–Neto, F. (2008b). On beta regression residuals. *Journal of Applied Statistics*, **35**, 407–419.

19. Ferrari, S.L.P. & Cribari–Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815.

20. Ferrari, S.L.P., Espinheira, P.L. & Cribari–Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Sinica*, **65**, 337–351.

21. Ferrari, S.L.P. & Pinheiro, E.C. (2012). Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation*, **81**, 431–443.

22. Figueroa–Zúñiga, Arellano–Valle, R.B. & Ferrari, S.L.P. (2013). Mixed beta regression: A Bayesian Perspective. *Computational Statistics and Data Analysis*. DOI: 10.1016/j.csda.2012.12.002.

23. Guolo, A. & Varin, C. (2012). Marginal beta regression for bounded time series. Technical report. Available at http://www.dst.unive.it/∼sammy/Home_files/guolo_varin_techreport.pdf.

24. Kieschnick, R. & McCullough, B.D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3, 193–213.

25. Kosmidis, I. & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, **4**, 1097–1112.

26. Ospina, R., Cribari–Neto, F. & Vasconcellos, K.L.P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis*, **51**, 960–981. Erratum at *Computational Statistics and Data Analysis*, **55**, 2445.

27. Ospina, R. & Ferrari, S.L.P. (2010). Inflated beta distributions. *Statistical Papers*, **51**, 111–126.

28. Ospina, R. & Ferrari, S.L.P. (2012a). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis*, **56**, 1609–1623.

29. Ospina, R. & Ferrari, S.L.P. (2012b). On bias correction in a class of inflated beta regression models. *International Journal of Statistics and Probability*, **1**. doi:10.5539/ijsp.v1n2p269.

30. Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9, 325–346.

31. Pereira, G.H.A., Botter, D.A. & Sandoval, M.C. (2012). The truncated inflated beta distribution. *Communications in Statistics - Theory and Methods*, **41**, 907–919.

32. Pereira, G.H.A., Botter, D.A. & Sandoval, M.C. (2013). A regression model for special proportions. *Statistical Modeling*. To appear.

33. Pereira, T.L.& Cribari–Neto, F. (2013). Detecting model misspecification in inflated beta regressions. *Communications in Statistics – Simulation and Computation*. To appear.

34. Pereira, T.L. ; Cribari–Neto, F. (2012). Modified likelihood ratio statistics for inflated beta regressions. *Journal of Statistical Computation and Simulation*. DOI:10.1080/00949655.2012.736514.

35. Ramalho E.A., Ramalho, J.J.S., Murteira, J.M.R. (2010). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, **25**, 19–68.

36. Rocha, A.V. & Cribari–Neto (2009). Beta autoregressive moving average models. *Test*, **18**, 529–545.

37. Rocha, A.V. & Simas, A.B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, **20**, 95–119.

38. Rydlewski, J.P. (2007). Beta-regression model for periodic data with a trend. *Universitatis Iagellonicae Acta Mathematica*, **XLV**, 211–222.

39. Rydlewski, J.P. & Mielczarek, D. (2012). On the maximum likelihood estimator in the geeralized beta regression model. *Opuscula Mathematica*, **32**, 761–774.

40. Smithson, M. & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methdos*, 11, 54–71.

41. Souza, D.F. & Moura, F.A.S. (2012a). Multivariate beta regression. *Technical Report*. Available at http://www.dme.im.ufrj.br/arquivos/publicacoes/arquivo245.pdf.

42. Souza, D.F. & Moura, F.A.S. (2012b). Multivariate beta regression with application to small area estimation. *Technical Report*. Available at http://www.dme.im.ufrj.br/arquivos/publicacoes/arquivo246.pdf.

43. Verkuilen, J. & Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, **37**, 82–113.

44. Weihua, Z., Riquan, Z., Zhensheng, H. & Jingyan, F. (2012). Partially linear single-index beta regression model and score test. *Journal of Multivariate Analysis*, 103, 16–123.

45. Wu, Y., Fedorov, V.V. & Propert, K.J. (2005). Optimal design for dose response using beta distributed responses. *Journal of Biopharmaceutical Statistics*, **15**, 753–771.

46. Zimprich, D. (2010). Modeling change in skewed variables using mixed beta regression models. *Research in Human Development*, **7**, 9–26.

Applications

1. Abdo, J.A., Cirano, F.R., Casati, M.Z., Ribeiro, F.V., Giampaoli, V., Casarin, R.C.V., Pimentel, S.P. (2012). Influence of Dyslipidemia and Diabetes Mellitus on Chronic Periodontal Disease. *Journal of Periodontology*. DOI: 10.1902/jop.2012.120366.
2. Balia, S. (2011). Survival expectations, subjective health and smoking: evidence from European countries. Working paper HEDG 11/30, University of York. Available at http://www.york.ac.uk/res/herc/documents/wp/11_30.pdf.
3. Berggren, N., Daunfeldt, S.–O. & Hellström, J. (2012). Social trust and central-bank independence. IFN Working Paper 920. Available at http://dx.doi.org/10.2139/ssrn.2065320.
4. Buntaine, M.T. (2011). Does the Asian Development Bank Respond to Past Environmental Performance when Allocating Environmentally Risky Financing. *World Development*, *39*, 336–350.
5. Carmichael, C.S. (2006). Modelling student performance in a tertiary preparatory course. Master Thesis. University of Southern Queensland. Available at http://eprints.usq.edu.au/3577.
6. Castellani, M., Pattitoni, P., Scorcu, A.E. (2012). Visual artist price heterogeneity. *Economics and Business Letters*, **1**, 16–22.
7. Chylack Jr, L.T., Peterson, L.E., Feiveson, A.H., Wear, M.L., Keith Manuel, F., Tung, W.H. Hardy, D.S., Marak, L.J. & Cucinotta, F.A. (2009). NASA study of cataract in astronauts (NASCA). Report 1: Cross-sectional study of the relationship of exposure to space radiation and risk of lens opacity. *Radiation Research*, **172**, 10–20.
8. De Paola, M., Scoppa, V., Lombardo, R. (2010). Can gender quotas break down negative stereotypes? Evidence from changes in electoral rules. *Journal of Public Economics*, **94**, 344–353.
9. Egleston, B., Scharfstein, D.O., Munoz, B. & West, S. (2006). Investigating mediation when counterfactuals are not metaphysical: does sunlight UVB exposure mediate the effects of eyeglasses on cataracts? Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 113. http://biostats.bepress.com/jhubiostat/paper113.
10. Ekleston, N.I., Madsen, L., Hagar, J.C. & Temesgen, H. (2011). Estimating riparian understory vegetation cover with beta regression and copula models. *Forest Science*, **57**, 212–221.
11. Gallardo, A., Bovea, M.D., Colomer, F.J., Prades, M. (2012). Analysis of collection systems for sorted household waste in Spain. *Waste Management*, **32**, 1623–1633.
12. Gillion, D.Q. (2008). Knocking on the president's door: changing the way we understand presidential responsiveness. Available at http://www.rochester.edu/college/psc/apwg/GillionAPWGFall08.pdf.
13. Huang,X. & Oosterlee, C.W. (2011). Generalized beta regression models for random loss given default. *The Journal of Credit Risk*, **7**, 45–70.
14. Hubben, G.A.A., Bishai, D., Pechlivanoglou, P., Cattelan, A.M., Grisetti, R., Facchin, C., Compostella, F.A., Bos, J.M., Postma, M.J, & Tramarin, A. The societal burden of HIV/AIDS in Northern Italy: An analysis of costs and quality of life. *AIDS Care*, **20**, 449–455.

15. Hunger, M., Doring, A., Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life score over time. *BMC Medical Research Methodology*, **12**, 144.

16. Ibáñez, M.V., Prades, M., Simó, A. (2011). Modelling municipal waste separation rates using generalized linear models and beta regression. *Resources, Conservation and Recycling*, **55**, 1129–1138.

17. Korhonen, L., Korhonen, K.T., Stenberg, P., Maltamo, M. & Rautiainen, M. (2007). Local models for forest canopy cover with beta regression. *Silva Fennica*, **41**, 671–685.

18. Ma, Y., Roy, J., Marcus, B. (2010). Causal models for randomized trials with two active treatments and continuous compliance. *Statistics in Medicine*, **30**, 2349–2362.

19. Peplonska, B., Bukowska, A., Sobala, W., Reszka, E., Gromadzinska, J., Wasowicz, W., Lie, J.A., Kjuus, H., Ursin, G. (2012). Rotating night shift work and mammographic density. *Cancer Epidemiology, Biomarkers and Prevention*. DOI: 10.1158/1055-9965.EPI-12-0005.

20. Pinto, E.R, Resende, L.A., Pereira, L.O., Destro Filho, J.B., (2011). Modelos estatísticos para estimação da área miocárdica sob risco de necrose. *Revista Brasileira de Biometria*, **29**, 395–415.

21. Quintero, H.E., Durland, E., Davis, D.A. & Dumham, R. (2011). Effect of lipid supplementation on reproductive performance of female channel catfish, Ictalurus punctatus, induced and strip-spawned for hybridization. **17**, 117–129.

22. Rogers, J.A., Polhamus, D., Gillespie, W.R., Ito, K., Romero, K., Qiu, R., Stephenson, D., Gastonguay, M.R., Corrigan, B. (2012). Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *Journal of Pharmacokinetics and Pharmacodynamics*, **39**, 479–498.

23. Wallis, E., Nally, R.M. & Lake, S. (2009). Do tributaries affect loads and fluxes of particulate organic matter, inorganic sediment and wood? Patterns in an upland river basin in south-eastern Australia. *Hydrobiologia*, **636**, 307–317.

Computational implementation

1. Cribari–Neto, F. & Zeiles, A. (2010). Beta regression in R. *Journal of Statistical Software*, **34**, 1–24.
2. Grün, B, Kosmidis, I. & Zeiles, A. (2012). Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*, **48**,11. Available at http://www.jstatsoft.org/v48/i11.
3. Stasinopoulos, D.M. & Rigby, R.A., (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 7, Available at http://www.jstatsoft.org/v23/i07/paper.
4. Swearingen, C.J., Castro, M.S.M. & Bursac, Z. (2011). Modeling percentage outcomes: the %Beta_Regression macro. *SAS Global Forum 2011*, Paper 335-2011. Available at http://support.sas.com/resources/papers/proceedings11/335-2011.pdf.
5. Swearingen, C.J., Castro, M.S.M. & Bursac, Z. (2012). Inflated beta regression: zero, one, and everything in between. *SAS Global Forum 2012*, Paper 325-2012. Available at http://support.sas.com/resources/papers/proceedings12/325-2012.pdf