

# A Short Course in Beta Regression Models

Raydonal Ospina Martínez

Department of Statistics  
Federal University of Pernambuco  
Cidade Universitária  
Recife/PE 50740-540, Brazil

E-mail: [raydonal@ufpe.br](mailto:raydonal@ufpe.br)  
URL: <http://www.de.ufpe.br/~raydonal/>

XXI SIMPOSIO DE ESTADÍSTICA “MODELOS DE REGRESIÓN”  
19 al 23 de julio de 2011

This material is based mainly on the papers [Cribari-Neto, F., and Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2), 1-24.] and [Simas, A.B., and Barreto-Souza, W., and Rocha, A.V. (2010). Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, 54(2), 348-366.]. Some snippets of text in the document are literal copies of the papers.

## Outline of the Course

- ▶ Two lectures, each 1.5 hours.
- ▶ Lecture 1. Basics in Beta Regression Models.
  - ▶ Motivation.
  - ▶ Model.
  - ▶ Estimation.
  - ▶ Testing.
  - ▶ Diagnostics.
- ▶ Lecture 2. Beta regression models in practice.
  - ▶ Applications.

# Lecture 1. Basics in Beta Regression Models.

## Motivation

How should we one perform a regression analysis in which the dependent variable is restricted to the standard unit interval such as rates and proportions?

## Possible solution

- ▶ Transform the dependent variable  $y$  so that it assumes values on the real line. (Example:  $\tilde{y} = \log(y/(1 - y))$ .)
- ▶ This approach, however, has drawbacks, one of them being the fact that the model parameters cannot be easily interpreted in terms of the original response.
- ▶ Regressions involving data from the unit interval are typically heteroskedastic: they display more variation around the mean and less variation as we approach the lower and upper limits of the standard unit interval.
- ▶ Another shortcoming is that measures of proportions typically display asymmetry, and hence inference based on the normality assumption can be misleading.

## An intelligent approach

- ▶ [Ferrari and Cribari-Neto, 2004] proposed a regression model for continuous variates that assume values in the standard unit interval, e.g., rates, proportions, or concentrations indices.
- ▶ The model is based on the assumption that the response is beta-distributed, they called their model *the beta regression model*.
- ▶ The regression parameters are interpretable in terms of the mean of  $y$  (the variable of interest) and the model is naturally heteroskedastic and easily accommodates asymmetries.
- ▶ A variant of the beta regression model that allows for nonlinearities and variable dispersion was proposed by [Simas et al., 2010].

- ▶ The chief motivation for the beta regression model lies in the flexibility delivered by the assumed beta law.
- ▶ The beta density can assume a number of different shapes depending on the combination of parameter values, including left- and right-skewed or the flat shape of the uniform density (which is a special case of the more general beta density).
- ▶ The beta density is usually expressed as

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, \quad 0 < y < 1,$$

where  $p, q > 0$  and  $\Gamma(\cdot)$  is the gamma function.

- ▶ A beta regression model based on this parameterization was proposed by [Vasconcellos and Cribari-Neto, 2005].



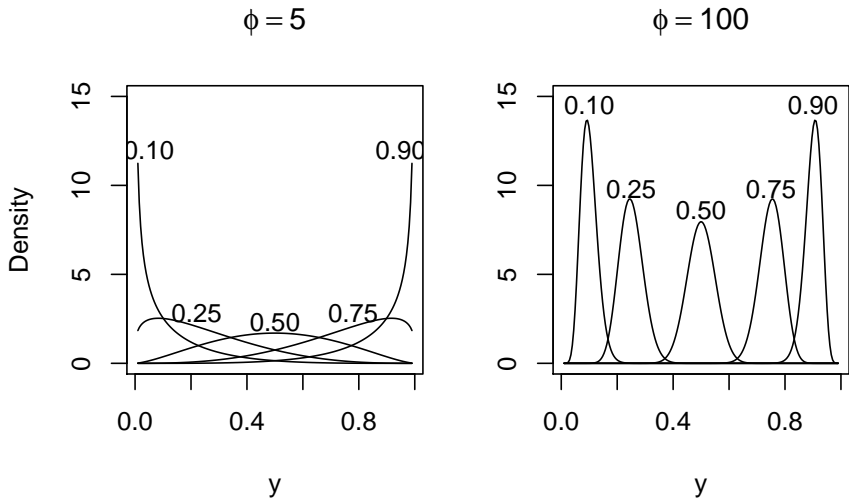
## Density

- ▶ [Ferrari and Cribari-Neto, 2004] proposed a different parameterization by setting  $\mu = p/(p + q)$  and  $\phi = p + q$ :

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

with  $0 < \mu < 1$  and  $\phi > 0$ .

- ▶ We write  $y \sim \mathcal{B}(\mu, \phi)$ . Here,  $E(y) = \mu$  and  $\text{VAR}(y) = \mu(1 - \mu)/(1 + \phi)$ .
- ▶ The parameter  $\phi$  is known as the precision parameter since, for fixed  $\mu$ , the larger  $\phi$  the smaller the variance of  $y$ ;  $\phi^{-1}$  is a dispersion parameter.



**Figure 1:** Probability density functions for beta distributions with varying parameters  $\mu = 0.10, 0.25, 0.50, 0.75, 0.90$  and  $\phi = 5$  (left) and  $\phi = 100$  (right).

## The model - General version -

- ▶ Let  $y = (y_1, \dots, y_n)^T$  be a random sample, where  $y_i \sim \mathcal{B}(\mu_i, \phi_i)$ ,  $i = 1, \dots, n$ .
- ▶ Suppose the mean and the precision parameter of  $y_i$  satisfies the following functional relations:

$$\begin{aligned}g_1(\mu_i) &= \eta_{1i} = f_1(x_i^T; \beta) \\g_2(\phi_i) &= \eta_{2i} = f_2(z_i^T; \theta).\end{aligned}\tag{1}$$

## The model

- ▶ Here,  $\beta = (\beta_1, \dots, \beta_k)^T$  and  $\theta = (\theta_1, \dots, \theta_h)^T$  are vectors of unknown regression parameters which are assumed to be functionally independent,  $\beta \in \mathbb{R}^k$  and  $\theta \in \mathbb{R}^h$ ,  $k + h < n$ .
- ▶  $\eta_{1i}$  and  $\eta_{2i}$  are predictors.
- ▶  $x_{i1}, \dots, x_{iq_1}, z_{i1}, \dots, z_{iq_2}$  are observations on  $q_1$  and  $q_2$  *known* covariates, which need not to be exclusive.
- ▶ We assume that the derivative matrices  $\tilde{X} = \partial\eta_1/\partial\beta$  and  $\tilde{Z} = \partial\eta_2/\partial\theta$  have rank  $k$  and  $h$ , respectively.
- ▶  $\phi^{-1}$  is a dispersion parameter.

## The model

- ▶ The link functions  $g_1 : (0, 1) \rightarrow \mathbb{R}$  and  $g_2 : (0, \infty) \rightarrow \mathbb{R}$  are strictly monotonic and twice differentiable.
- ▶ A number of different link functions can be used, such as the logit specification  $g_1(\mu) = \log\{\mu/(1 - \mu)\}$ , the probit function  $g_1(\mu) = \Phi^{-1}(\mu)$ , where  $\Phi(\cdot)$  denotes the standard normal distribution function, the complementary log-log function  $g_1(\mu) = \log\{-\log(1 - \mu)\}$ , among others, and for  $g_2$ ,  $g_2(\phi) = \log \phi$ , the logarithmic function,  $g_2(\phi) = \sqrt{\phi}$ , the square root function,  $g_2(\phi) = \phi$  (with special attention on the positivity of the estimates), among others.
- ▶ A rich discussion of link functions can be found in McCullagh and Nelder (1989); see also Atkinson (1985, Chapter 7).

## Especial cases

- ▶ The linear beta regression model

We have, in (1),  $g_1(\mu_i) = g(\mu_i) = \eta_{1i} = x_i^T \beta$ , where  $g(\cdot)$  is some link function and  $g_2(\phi) = \phi_i = \phi$ . We have that in this case  $\tilde{X} = X$  and  $\tilde{Z} = \mathbf{1}$  where  $X$  is the matrix of covariates with rows given by  $x_i^T$ .

- ▶ The linear beta regression model with dispersion covariates

The precision parameter  $\phi$  vary through a linear regression structure. More precisely, in (1),  $g_1(\mu_i) = g(\mu_i) = \eta_{1i} = x_i^T \beta$ , and  $g_2(\phi_i) = \eta_{2i} = z_i^T \theta$ . In this case  $\tilde{X} = X$  and  $\tilde{Z} = Z$  where  $X$  and  $Z$  are covariates matrix with rows given by  $x_i^T$  and  $z_i^T$ , respectively.

## Especial cases

- ▶ The nonlinear beta regression model with linear dispersion covariates

The precision parameter  $\phi$  vary through a linear regression structure. For this model the equation (1) becomes

$$g_1(\mu_i) = \eta_{1i} = f(x_i^T; \beta) \quad \text{and} \quad g_2(\phi_i) = \eta_{2i} = z_i^T \theta,$$

where  $\beta \in \mathbb{R}^k$  and  $\theta \in \mathbb{R}^h$ . Then, we have that for this model  $\tilde{X}$  remaining the same, and  $\tilde{Z} = Z$ , where  $Z$  is the matrix of covariates with rows given by  $z_i^T$ .

## The log-likelihood function

- ▶ For this class of beta regression models has

$$\ell(\beta, \theta) = \sum_{i=1}^n \ell_i(\mu_i, \phi_i), \quad (2)$$

where

$$\begin{aligned} \ell_i(\mu_i, \phi_i) &= \log \Gamma(\phi_i) - \log \Gamma((1 - \mu_i)\phi_i) + (\mu_i\phi_i - 1) \log y_i \\ &\quad + \{(1 - \mu_i)\phi_i - 1\} \log(1 - y_i); \end{aligned}$$

$\mu_i = g_1^{-1}(\eta_{1i})$ ,  $\phi_i = g_2^{-1}(\eta_{2i})$ , as defined in (1), are functions of  $\beta$  and  $\theta$ , respectively.

- ▶ It is possible to show that this beta regression model is regular.



## Score function for $\beta$ 's

- ▶ The components of the score vector, obtained by differentiation of the log-likelihood function with respect to the parameters.
- ▶ For  $r = 1, \dots, k$ , as

$$U_r(\beta, \theta) = \frac{\partial \ell(\beta, \theta)}{\partial \beta_r} = \sum_{i=1}^n \phi_i(y_i^* - \mu_i^*) \frac{d\mu_i}{d\eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_r},$$

where  $d\mu_i/d\eta_{1i} = 1/g'_1(\mu_i)$ ,  $y_i^* = \log(y_i/(1 - y_i))$ ,  $\mu_i^* = \psi(\mu_i\phi_i) - \psi((1 - \mu_i)\phi_i)$ , and  $\psi(\cdot)$  is the digamma<sup>1</sup> function.

---

<sup>1</sup>We denote generally the polygamma function by  $\psi^{(m)}(\cdot)$ ,  $m = 0, 1, \dots$ , where  $\psi^{(m)}(x) = (d^{m+1}/dx^{m+1}) \log \Gamma(x)$ ,  $x > 0$ .

## Score function for $\theta$ 's

- ▶ For

$$U_R(\beta, \theta) = \frac{\partial \ell(\beta, \theta)}{\partial \theta_R}$$
$$= \sum_{i=1}^n \{ \mu_i (y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1 - \mu_i)\phi_i) + \log(1 - y_i) \} \frac{d\phi_i}{d\eta_{2i}} \frac{\partial \eta_{2i}}{\partial \theta_R},$$

where  $d\phi_i/d\eta_{2i} = 1/g'_2(\phi_i)$ , and  $R = 1, \dots, h$ .

- ▶ Further, the regularity conditions implies that

$$E \left( \log \frac{y_i}{1 - y_i} \right) = \psi(\mu_i \phi_i) - \psi((1 - \mu_i)\phi_i),$$

and

$$E\{\log(1 - y_i)\} = \psi((1 - \mu_i)\phi_i) - \psi(\phi_i).$$

## Score vector

- ▶ Consider the complete parameter vector  $\zeta = (\beta^T, \theta^T)^T$ .
- ▶ Define the vectors  $y^* = (y_1^*, \dots, y_n^*)^T$ ,  $\mu^* = (\mu_1^*, \dots, \mu_n^*)^T$ ,  $v = (v_1, \dots, v_n)^T$ , where  
 $v_i = \mu_i(y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1 - \mu_i)\phi_i) + \log(1 - y_i)$ .
- ▶ The matrix  $T_1 = \text{diag}(d\mu_i/d\eta_{1i})$ ,  $T_2 = \text{diag}(d\phi_i/d\eta_{2i})$ ,  $\Phi = \text{diag}(\phi_i)$ .
- ▶ The  $(k + h) \times 1$  dimensional score vector  $U(\zeta)$  in the form  $(U_\beta(\beta, \theta)^T, U_\theta(\beta, \theta)^T)^T$ , with

$$\begin{aligned}U_\beta(\beta, \theta) &= \tilde{X}^T \Phi T_1 (y^* - \mu^*), \\U_\theta(\beta, \theta) &= \tilde{Z}^T T_2 v.\end{aligned}\tag{3}$$

## Fisher's information matrix

- ▶ It is possible to obtain Fisher's information matrix for the parameter vector  $\zeta = (\beta^T, \theta^T)^T$  as

$$K(\zeta) = P^T W P.$$

- ▶ Define  $P$  as the  $2n \times (k + h)$  dimensional matrix

$$P = \begin{pmatrix} \tilde{X} & 0 \\ 0 & \tilde{Z} \end{pmatrix}. \quad (4)$$

- ▶ let  $W$  be the  $2n \times 2n$  matrix

$$W = \begin{pmatrix} W_{\beta\beta} & W_{\beta\theta} \\ W_{\beta\theta} & W_{\theta\theta} \end{pmatrix}, \quad (5)$$

► Here

$$W_{\beta\beta} = \text{diag} \left( \phi_i^2 a_i \left( \frac{d\mu_i}{d\eta_{1i}} \right)^2 \right),$$

$$W_{\beta\theta} = \text{diag} \left( \phi_i \{ \mu_i a_i - \psi'((1 - \mu_i)\phi_i) \} \left( \frac{d\mu_i}{d\eta_{1i}} \right) \left( \frac{d\phi_i}{d\eta_{2i}} \right) \right),$$

$$W_{\theta\theta} = \text{diag} \left( b_i \left( \frac{d\phi_i}{d\eta_{2i}} \right)^2 \right).$$

► where,  $a_i = \psi'((1 - \mu_i)\phi_i) + \psi'(\mu_i\phi_i)$  and  
 $b_i = \psi'((1 - \mu_i)\phi_i)(1 - \mu_i)^2 + \psi'(\mu_i\phi_i)\mu_i^2 - \psi'(\phi_i)$ .

## Estimation

- ▶ Note that  $W_{\beta\theta} \neq 0$ , thus indicating that the parameters  $\beta$  and  $\theta$  are not orthogonal, in contrast to the class of generalized linear models (McCullagh and Nelder, 1989), where such orthogonality holds.
- ▶ Nevertheless, the MLEs  $\hat{\zeta}$  and  $K(\hat{\zeta})$  are consistent estimators of  $\zeta$  and  $K(\zeta)$ , respectively, where  $K(\hat{\zeta})$  is the Fisher's information matrix evaluated at  $\hat{\zeta}$ .
- ▶ The MLEs of  $\beta$  and  $\theta$  are obtained as the solution of the nonlinear system  $U(\zeta) = 0$ . In practice, the MLEs can be obtained through a numerical maximization of the log-likelihood function using a nonlinear optimization algorithm, e.g., BFGS. For details, see Press et al. (1992).

## Estimation

- ▶ As initial guesses for  $\beta$  and  $\theta$ , we suggest one to obtain the estimates from the following normal nonlinear regression model with dispersion covariates:  $g_1(\mu_i) = f_1(x_i; \beta)$  and  $g_2(\sigma_i^2) = f_1(z_i; \theta)$ .
- ▶ This will produce  $\hat{\beta}^{(0)}$  and  $\hat{\theta}^{(0)}$ , which will be our initial guesses, where for this initial guess we assume that  $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ .
- ▶ Note that we are viewing the normal nonlinear model above in the generalized nonlinear model sense, since we are using the link functions.
- ▶ The R package `nlnme2` fits such a model.

- ▶ The beta distribution has been incorporated in the GAMLSS framework (Rigby & Stasinopoulos, 2005).
- ▶ GAMLSS allows the flexible modeling of each of the three parameters that index the distribution using parametric terms involving linear or nonlinear predictors, smooth nonparametric terms (e.g., cubic splines or loess), and random effects.
- ▶ Maximum (penalized) likelihood estimation is approached through a Newton-Raphson or Fisher scoring algorithm with the backfitting algorithm for the additive components.



- ▶ The GAMLSS approach consists of an application of the `gamlss` functions, which are fully documented in the `gamlss` package (Stasinopoulos & Rigby, 2007; see also <http://www.gamlss.org>).
- ▶ The structure of the `gamlss` functions is familiar to readers who are used to the R (or S-Plus) syntax (the `glm` function, in particular).
- ▶ The set of `gamlss` packages can be freely downloaded from the R library at <http://www.r-project.org/>.

## In large samples

- ▶ If  $J(\zeta) = \lim_{n \rightarrow \infty} K(\zeta)/n$  exists and is nonsingular, we have that

$$\sqrt{n} (\hat{\zeta} - \zeta) \xrightarrow{d} N_{k+h}(0, J(\zeta)^{-1}).$$

- ▶ If  $\zeta_r$  denotes the  $r$ th component of  $\zeta$ , it follows that

$$(\hat{\zeta} - \zeta) \{K(\hat{\zeta})^{rr}\}^{-1/2} \xrightarrow{d} N(0, 1),$$

where  $K(\hat{\zeta})^{rr}$  is the  $r$ th diagonal element of  $K(\hat{\zeta})^{-1}$ .

- ▶ For  $0 < \alpha < 1/2$ , and  $q_\gamma$  representing the  $\gamma$  quantile of the  $N(0, 1)$  distribution, we have, for  $r = 1, \dots, k$ ,

$$\hat{\beta}_r \pm q_{1-\alpha/2} \left( K(\hat{\zeta})^{rr} \right)^{1/2}$$

as the limits of asymptotic confidence intervals for  $\beta_r$  with asymptotic coverage of  $100(1 - \alpha)\%$ .

## Testing

- ▶ The likelihood ratio, Rao's score, and Wald's (W) statistics to test hypotheses on the parameters can be calculated from the log-likelihood function, the score vector, the Fisher information matrix, and its inverse given above.
- ▶ Their null distributions are usually unknown and the tests rely on asymptotic approximations.
- ▶ In large samples, a chi-squared distribution can be used as an approximation to the true null distributions.

- ▶ For testing the significance of the  $i$ th regression parameter that models  $\mu$ , one can use the signed square root of Wald's statistic,  $\widehat{\beta}_i / \text{s.e.}(\widehat{\beta}_i)$ , where  $\text{s.e.}(\widehat{\beta}_i)$  is the asymptotic standard error of the MLE of  $\beta_i$  obtained from the inverse of Fisher's information matrix evaluated at the maximum likelihood estimates.
- ▶ The limiting null distribution of the test statistic is standard normal.
- ▶ Significance tests on the  $\theta$ 's can be performed in a similar fashion.

- ▶ A RESET-type misspecification test for fixed dispersion beta regression, similar to that of [Ramsey, 1969] for linear regressions, was proposed by [Cribari-Neto and Lima, 2007].
- ▶ The authors considered different variants of the test and concluded that the best performing testing strategy in finite samples is to include  $\hat{\eta}_i^2 = (x_i^\top \hat{\beta})^2$ , where  $\hat{\beta}$  denotes the ML estimator of  $\beta$ , as an additional regressors in an augmented (artificial) regression, and then test its exclusion using a score test.
- ▶ Rejection of the null hypothesis suggests that the model is misspecified. Misspecification can follow, e.g., from incorrectly specifying the link function, from neglecting nonlinearities or from omitting important regressors.

## Model selection

- ▶ Nested beta regression models can be compared via the likelihood ratio test, using twice the difference between the maximized log-likelihoods of a full model and a restricted model whose covariates are a subset of the full model.
- ▶ Information criteria, such as the generalized Akaike information criterion (GAIC),  $GAIC = \hat{D} + d_{\emptyset}$ , can be used for comparing non-nested models.
- ▶  $\hat{D} = -2\hat{\ell}$  is the global fitted deviance (Rigby & Stasinopoulos, 2005),  $\hat{\ell}$  is the maximized log-likelihood and  $d$  is the dimension of  $\zeta$ .
- ▶ The model with the smallest GAIC is then selected.

## Diagnostics - Generalized leverage -

- ▶ The generalized leverage (GL) proposed by Wei et al. (1998) is defined by

$$GL(\tilde{\zeta}) = \frac{\partial \tilde{y}}{\partial y^T}$$

where  $\zeta$  is an  $s$ -vector such that  $E(y) = \mu(\zeta)$  and  $\tilde{\zeta}$  is an estimator of  $\zeta$ , with  $\tilde{y} = \mu(\tilde{\zeta})$ .

- ▶ The  $(i, l)$  element of  $GL(\tilde{\zeta})$  i.e. the GL of the estimator  $\tilde{\zeta}$  at  $(i, l)$ , is the instantaneous rate of change in the  $i$ th predicted value with respect to the  $l$ th response value.
- ▶ The GL is obtained as

$$D_{\zeta} \left( -\frac{\partial^2 \ell}{\partial \zeta \partial \zeta^T} \right)^{-1} \frac{\partial^2 \ell}{\partial \zeta \partial y^T} \quad (6)$$

evaluated at  $\hat{\zeta}$ , where  $D_{\zeta} = \partial \mu / \partial \zeta^T$ .

## Diagnostics - Residuals -

- ▶ The raw response residuals  $y_i - \hat{\mu}_i$  are typically not used due to the heteroskedasticity inherent in the model.
- ▶ A natural alternative are *Pearson residuals* which [Ferrari and Cribari-Neto, 2004] call *standardized ordinary residuals* and define as

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{VAR}}(y_i)}}, \quad (7)$$

where  $\widehat{\text{VAR}}(y_i) = \hat{\mu}_i(1 - \hat{\mu}_i)/(1 + \hat{\phi}_i)$ ,  $\hat{\mu}_i = g_1^{-1}(f_1(x_i^\top; \hat{\beta}))$ , and  $\hat{\phi}_i = g_2^{-1}(f_2(z_i^\top; \hat{\gamma}))$ .

- ▶ Deviance residuals can be defined in the standard way via signed contributions to the excess likelihood.



- ▶ Further residuals were proposed by [Espinheira et al., 2008b], in particular one residual with better properties that they named *standardized weighted residual 2 our score residual*:

$$r_{\text{sw2},i} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\hat{v}_i(1 - h_{ii})}}, \quad (8)$$

where  $y_i^* = \log\{y_i/(1 - y_i)\}$  and  $\mu_i^* = \psi(\mu_i\phi) - \psi((1 - \mu_i)\phi)$ ,  $\psi(\cdot)$  denoting the digamma function. Standardization is then by  $v_i = \{\psi'(\mu_i\phi) + \psi'((1 - \mu_i)\phi)\}$  and  $h_{ii}$ , the  $i$ th diagonal element of the hat matrix.

- ▶ The problem with the above residual is that it does not take into account the discrepant values in the covariate matrix  $Z$ .

- ▶ [Rocha and Simas, 2010] proposed a modification of the score residual that takes into account all the discrepancy of the model.

$$r_{MS,i} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\hat{v}_i(1 - \hat{G}L_{ii})}}, \quad (9)$$

where  $GL_{ii}$  is the  $i$ th element of the diagonal of the generalized leverage matrix given in (6).

- ▶ Another alternative. Use the randomized quantile residual (Dunn & Smyth, 1996). It is a randomized version of the Cox & Snell (1968) residual and given by

$$r_t^q = \Phi^{-1}(u_t), \quad t = 1, \dots, n, \quad (10)$$

where  $\Phi(\cdot)$  denotes the standard normal distribution function.

## Diagnostics -Envelope plot -

- ▶ A plot of these residuals against the index of the observations ( $i$ ) should show no detectable pattern.
- ▶ A detectable trend in the plot of some residual against the predictors may be suggestive of link function misspecification.
- ▶ Normal probability plots with simulated envelopes are a helpful diagnostic tool (Atkinson, 1985).
- ▶ Simulation results not presented here indicated that the residuals perform well in detecting whether the distribution assumption is incorrect.

## Diagnostics - Global goodness-of-fit measure -

- ▶ A simple global goodness-of-fit measure is a pseudo  $R^2$ , say  $R_p^2$  defined by the square of the sample correlation coefficient between the outcomes,  $y_1, \dots, y_n$ , and their corresponding predicted values,  $\widehat{\mu}_1, \dots, \widehat{\mu}_n$ .
- ▶ A perfect agreement between the  $y$ 's and  $\widehat{\mu}$ 's yields  $R_p^2 = 1$ .
- ▶ Other pseudo  $R^2$ 's are defined as  $R_p^{2*} = 1 - \log \widehat{L} / \log \widehat{L}_0$  (McFadden, 1974) and  $R_{LR}^2 = 1 - (\widehat{L}_0 / \widehat{L})^{2/n}$  (Cox and Snell, 1989, p. 208-209), where  $\widehat{L}_0$  and  $\widehat{L}$  are the maximized likelihood functions of the null model and the fitted model, respectively.
- ▶ The ratio of the likelihoods or log-likelihoods may be regarded as measures of the improvement over the model with only three parameters  $\mu$  and  $\phi$ , achieved by the model under investigation.

## Diagnostics - Influence measures -

- ▶ A well-known measure of the influence of each observation on the regression parameter estimates is the likelihood displacement (Cook & Weisberg, 1982, Ch. 3).
- ▶ The likelihood displacement that results from removing the  $t$ th observation from the data is defined by

$$LD_t = 2\{\ell(\hat{\zeta}) - \ell(\hat{\zeta}_{(t)})\},$$

where  $d$  is the dimension of  $\zeta$  and  $\hat{\zeta}_{(t)}$  is the MLE of  $\zeta$  obtained after removing the  $t$ th observation from the data.

- ▶  $LD_t$  combines leverage and residuals.
- ▶ It is common practice to plot  $LD_t$  against  $t$ .
- ▶ Other diagnostic measures can be considered, such as local influence measures (Cook, 1986).

## Lecture 2. Beta regression models in practice.

## **betareg**: An implementation of beta regression models using R

- ▶ Beta regression as suggested by Ferrari and Cribari-Neto (2004) and extended by Simas, Barreto-Souza, and Rocha (2010) is implemented in the **betareg** package.
- ▶ It is modeled to be beta-distributed with parametrization using mean and precision parameter.
- ▶ The mean is linked, as in generalized linear models (GLMs), to the responses through a link function and a linear predictor.
- ▶ The precision parameter  $\phi$  can be linked to another (potentially overlapping) set of regressors through a second link function, resulting in a model with variable dispersion.

- ▶ Estimation is performed by maximum likelihood (ML) via `optim()` using analytical gradients and (by default) starting values from an auxiliary linear regression of the transformed response.
- ▶ The main model-fitting function in **betareg** is `betareg()` which takes a fairly standard approach for implementing ML regression models in R: `formula plus data` is used for model and data specification, then the likelihood and corresponding gradient (or estimating function) is set up, `optim()` is called for maximizing the likelihood, and finally an object of S3 class “betareg” is returned for which a large set of methods to standard generics is available.
- ▶ The workhorse function is `betareg.fit()` which provides the core computations without `formula`-related data pre-and post-processing.



- ▶ The arguments of `betareg()` are

```
betareg(formula, data, subset, na.action, weights,  
        offset, link = "logit", link.phi = NULL,  
        control = betareg.control(...), model =  
        TRUE, y = TRUE, x = FALSE, ...)
```

## Some methods for **betareg**

Function	Description
<code>print()</code> <code>summary()</code>	simple printed display with coefficient estimates standard regression output (coefficient estimates, standard errors, partial Wald tests); returns an object of class "summary.betareg" containing the relevant summary statistics (which has a <code>print()</code> method)
<code>coef()</code> <code>vcov()</code> <code>predict()</code>	extract coefficients of model (full, mean, or precision components), a single vector of all coefficients by default associated covariance matrix (with matching names) predictions (of means $\mu_i$ , linear predictors $\eta_{1i}$ , precision parameter $\phi_i$ , or variances $\mu_i(1 - \mu_i)/(1 + \phi_i)$ ) for new data
<code>fitted()</code> <code>residuals()</code>	fitted means for observed data extract residuals [Espinheira et al., 2008b, deviance, Pearson, response, or different weighted residuals, see], defaulting to standardized weighted residuals 2 from Equation 8
<code>estfun()</code>	compute empirical estimating functions (or score functions), evaluated at observed data and estimated parameters [Zeileis, 2006, see]
<code>bread()</code>	extract "bread" matrix for sandwich estimators [Zeileis, 2006, see]

Function	Description
<code>terms()</code> <code>model.matrix()</code> <code>model.frame()</code> <code>logLik()</code>	extract terms of model components extract model matrix of model components extract full original model frame extract fitted log-likelihood
<code>plot()</code> <code>hatvalues()</code> <code>cooks.distance()</code> <code>gleverage()</code>	diagnostic plots of residuals, predictions, leverages etc. hat values (diagonal of hat matrix) (approximation of) Cook's distance compute generalized leverage [Wei et al., 1998, Rocha and Simas, 2010]
<code>coefstest()</code> <code>waldtest()</code> <code>linear.hypothesis()</code> <code>lrtest()</code> <code>AIC()</code>	partial Wald tests of coefficients Wald tests of nested models Wald tests of linear hypotheses likelihood ratio tests of nested models compute information criteria (AIC, BIC, ...)

**Table 1:** Functions and methods for objects of class “betareg”. The first four blocks refer to methods, the last block contains generic functions whose default methods work because of the information supplied by the methods above.

# Beta regression in practice: An illustration

## Prater's gasoline yield data

- ▶ We estimate and compare various flavors of beta regression models for the gasoline yield data of [Prater, 1956]
- ▶ The variable of interest is `yield`, the proportion of crude oil converted to gasoline after distillation and fractionation, for which a beta regression model is rather natural.
- ▶ [Ferrari and Cribari-Neto, 2004] employ two explanatory variables: `temp`, the temperature (in degrees Fahrenheit) at which all gasoline has vaporized, and `batch`, a factor indicating ten unique batches of conditions in the experiments (depending on further variables).

# The basic model: Estimation, inference, diagnostics

[Ferrari and Cribari-Neto, 2004] start out with a model where `yield` depends on `batch` and `temp`, employing the standard logit or log-log link. In **betareg**, this can be fitted via

```
R> data("GasolineYield", package = "betareg")
R> gy_logit <- betareg(yield ~ batch + temp,
+ data = GasolineYield)

R> gy_loglog <- betareg(yield ~ batch + temp,
+ data = GasolineYield, link = "loglog")
```

```
R> summary(gy_logit)
```

```
Call:
```

```
betareg(formula = yield ~ batch + temp, data = GasolineYield)
```

```
Standardized weighted residuals 2:
```

```
      Min      1Q  Median      3Q      Max
-2.8750 -0.8149  0.1601  0.8384  2.0483
```

```
Coefficients (mean model with logit link):
```

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-6.1595710	0.1823247	-33.784	< 2e-16	***
batch1	1.7277289	0.1012294	17.067	< 2e-16	***
batch2	1.3225969	0.1179021	11.218	< 2e-16	***
batch3	1.5723099	0.1161045	13.542	< 2e-16	***
batch4	1.0597141	0.1023598	10.353	< 2e-16	***
batch5	1.1337518	0.1035232	10.952	< 2e-16	***
batch6	1.0401618	0.1060365	9.809	< 2e-16	***
batch7	0.5436922	0.1091275	4.982	6.29e-07	***
batch8	0.4959007	0.1089257	4.553	5.30e-06	***
batch9	0.3857929	0.1185933	3.253	0.00114	**
temp	0.0109669	0.0004126	26.577	< 2e-16	***

```
Phi coefficients (precision model with identity link):
```

	Estimate	Std. Error	z value	Pr(> z )	
(phi)	440.3	110.0	4.002	6.29e-05	***

```
---
```

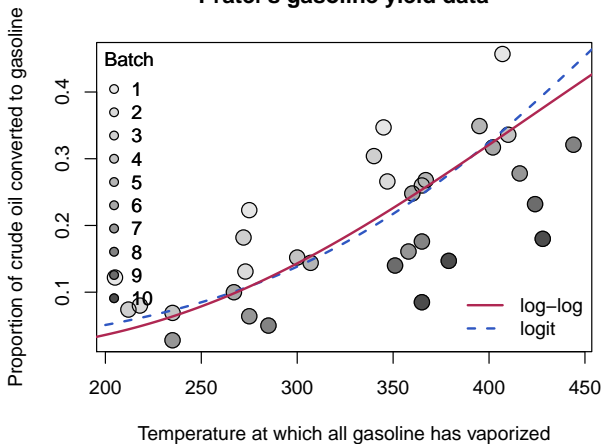
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Log-likelihood: 84.8 on 12 Df
```

```
Pseudo R-squared: 0.9617
```

```
Number of iterations in BFGS optimization: 51
```

## Prater's gasoline yield data



**Figure 2:** Gasoline yield data from [Prater, 1956]: Proportion of crude oil converted to gasoline explained by temperature (in degrees Fahrenheit) at which all gasoline has vaporized and given batch (indicated by gray level). Fitted curves correspond to beta regressions `gy_loglog` with log-log link (solid, red) and `gy_logit` with logit link (dashed, blue). Both curves were evaluated at varying temperature with the intercept for batch 6 (i.e., roughly the average intercept).

Goodness of fit is assessed using different types of diagnostic displays shown in their Figure 2. This graphic can be reproduced (in a slightly different order) using the `plot()` method for “betareg” objects, see Figure 3.

```
R> set.seed(123)
R> plot(gy_logit, which = 1:4, type = "pearson")
R> plot(gy_logit, which = 5, type = "deviance",
+ sub.caption = "")
R> plot(gy_logit, which = 1, type = "deviance",
+ sub.caption = "")
```



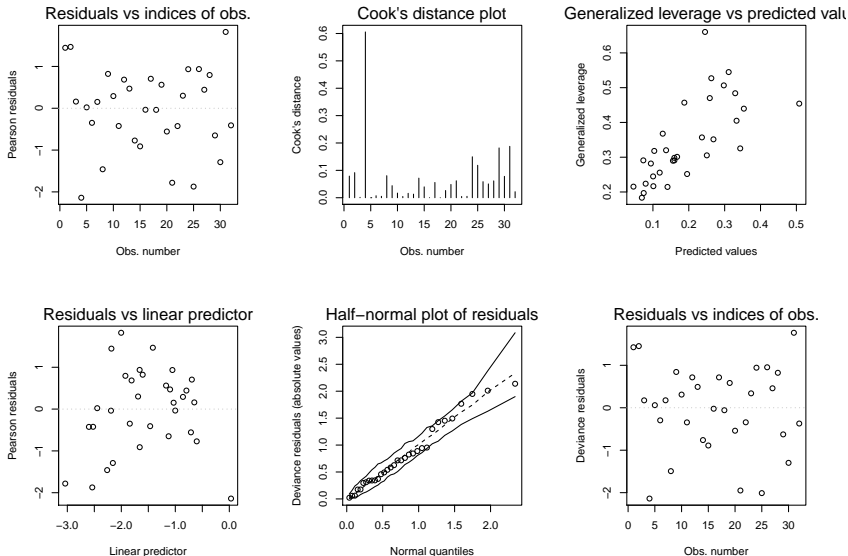


Figure 3: Diagnostic plots for beta regression model `gy_logit`.

As observation 4 corresponds to a large Cook's distance and large residual, [Ferrari and Cribari-Neto, 2004] decided to refit the model excluding this observation. While this does not change the coefficients in the mean model very much, the precision parameter  $\phi$  increases clearly.

```
R> gy_logit4 <- update(gy_logit, subset = -4)
R> coef(gy_logit, model = "precision")
```

```
(phi)
440.2783
```

```
R> coef(gy_logit4, model = "precision")
```

```
(phi)
577.7907
```

## Variable dispersion model

- ▶ The beta model already incorporates naturally a certain pattern in the variances of the response.
- ▶ It might be necessary to incorporate further regressors to account for heteroskedasticity [Simas et al., 2010].
- ▶ The Prater's gasoline yield data based on the same mean equation as above, but now with temperature  $t_{\text{temp}}$  as an additional regressor for the precision parameter  $\phi_j$ :

```
R> gy_logit2 <- betareg(yield ~ batch
+ + temp | temp, data = GasolineYield)
```

for which `summary(gy_logit2)` yields the MLE column in Table 19 of [Simas et al., 2010].

Here, only the parameters pertaining to  $\phi_i$  are reported

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.3641103  1.2257813  1.1128  0.2658
temp         0.0145703  0.0036183  4.0268 5.653e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

which signal a significant improvement by including the `temp` regressor.

Instead of using this Wald test, the models can also be compared by means of a likelihood-ratio test (see their Table 18) that confirms the results:

```
R> library("lmtest")  
R> lrtest(gy_logit, gy_logit2)
```

Likelihood ratio test

Model 1: yield ~ batch + temp

Model 2: yield ~ batch + temp | temp

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	12	84.798			
2	13	86.977	1	4.359	0.03681 *

---

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

Note that this can also be interpreted as testing the null hypothesis of equidispersion against a specific alternative of variable dispersion.

## Selection of different link functions

- ▶ Selection of an appropriate link function can greatly improve the model fit [McCullagh and Nelder, 1989],
- ▶ We replicate parts of the analysis in Section 5 of [Cribari-Neto and Lima, 2007].
- ▶ This reconsiders Prater's gasoline yield data but employs a log-log link instead of the previously used (default) logit link

```
R> gy_loglog <- betareg(yield ~ batch + temp,  
+ data = GasolineYield, link = "loglog")
```

- ▶ Clearly the pseudo  $R^2$  of the model is improved:

```
R> summary(gy_logit)$pseudo.r.squared
```

```
[1] 0.9617312
```

```
R> summary(gy_loglog)$pseudo.r.squared
```

```
[1] 0.9852334
```

- ▶ Similarly, the AIC<sup>2</sup> (and BIC) of the fitted model is not only superior to the logit model with fixed dispersion `gy_logit` but also to the logit model with variable dispersion `gy_logit2` considered in the previous section.

```
R> AIC(gy_logit, gy_logit2, gy_loglog)
```

	df	AIC
<code>gy_logit</code>	12	-145.5951
<code>gy_logit2</code>	13	-147.9541
<code>gy_loglog</code>	12	-168.3101

---

<sup>2</sup>Note that [Cribari-Neto and Lima, 2007] did not account for estimation of  $\phi$  in their degrees of freedom. Hence, their reported AICs differ by 2.

- ▶ If  $t_{emp}$  were included as a regressor in the precision equation of  $gy\_loglog$ , it would no longer yield significant improvements.
- ▶ Improvement of the model fit in the mean equation by adoption of the log-log link has waived the need for a variable precision equation.
- ▶ [Cribari-Neto and Lima, 2007] consider a sequence of diagnostic tests inspired by the RESET [Ramsey, 1969, regression specification error test] in linear regression models.
- ▶ To check for misspecifications, they consider powers of fitted means or linear predictors to be included as auxiliary regressors in the mean equation.
- ▶ In well-specified models, these should not yield significant improvements.



Analogous results can be obtained for `type = "response"` or higher powers.

```
R> lrtest(gy_logit, . ~ . + I(predict(gy_logit, type = "link")^2))
```

Likelihood ratio test

Model 1: yield ~ batch + temp

Model 2: yield ~ batch + temp + I(predict(gy\_logit, type = "link")^2)

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	12	84.798			
2	13	96.001	1	22.407	2.205e-06 ***

---

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

```
R> lrtest(gy_loglog, . ~ . + I(predict(gy_loglog, type = "link")^2))
```

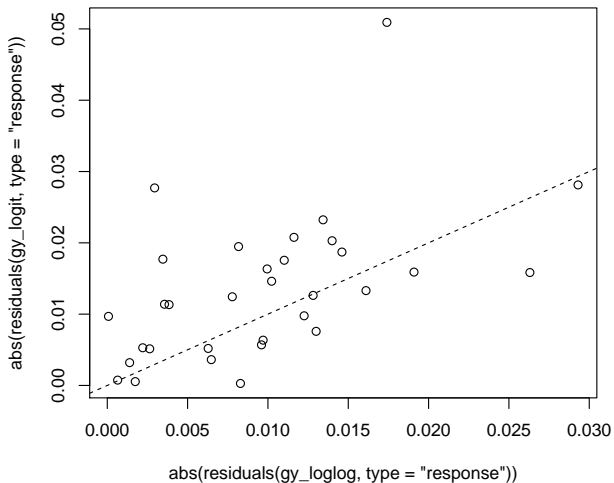
Likelihood ratio test

Model 1: yield ~ batch + temp

Model 2: yield ~ batch + temp + I(predict(gy\_loglog, type = "link")^2)

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	12	96.155			
2	13	96.989	1	1.6671	0.1966

- ▶ The improvement of the model fit can also be brought out graphically by comparing absolute raw residuals (i.e.,  $y_i - \hat{\mu}_i$ ) from both models.
- ▶ A different diagnostic display that is useful in this situation [Cribari-Neto and Lima, 2007, and is employed by] is a plot of predicted values ( $\hat{\mu}_i$ ) vs. observed values ( $y_i$ ) for each model. This can be created by `plot(gy_logit, which = 6)` and `plot(gy_loglog, which = 6)`, respectively.



**Figure 4:** Scatterplot comparing the absolute raw residuals from beta regression modes with log-log link (x-axis) and logit link (y-axis).

## Especial topics

- ▶ [Simas et al., 2010] further extend the beta regression model by allowing nonlinear predictors.
- ▶ Analytical bias corrections for the ML estimators of the parameters and generalizing is discussed by [Ospina et al., 2006] and [Simas et al., 2010].
- ▶ Diagnostics for the beta regression models is discussed by [Espinheira et al., 2008b], [Espinheira et al., 2008a], [Ferrari. et al., 2011], and [Rocha and Simas, 2010] among others.

- ▶ A general class of regression models for continuous proportions when the data contain zeros or ones is proposed by [Ospina and Ferrari, 2011].
- ▶ The proposed class of models assumes that the response variable has a mixed continuous-discrete distribution with probability mass at zero or one. The beta distribution is used to describe the continuous component of the model.
- ▶ A suitable parameterization of the beta law in terms of its mean and a precision parameter is used [Ospina and Ferrari, 2010].
- ▶ The parameters of the mixture distribution are modeled as functions of regression parameters.
- ▶ The authors provide inference, diagnostic, and model selection tools for this class of models.

# Bibliography I



Cribari-Neto, F. and Lima, L. B. (2007).  
A misspecification test for beta regressions.  
Technical report.



Espinheira, P. L., Ferrari, S. L. P., and Cribari-Neto, F. (2008a).  
Influence diagnostics in beta regression.  
*Computational Statistics & Data Analysis*, 52(9):4417–4431.



Espinheira, P. L., Ferrari, S. L. P., and Cribari-Neto, F. (2008b).  
On beta regression residuals.  
*Journal of Applied Statistics*, 35(4):407–419.



Ferrari, S. L. P. and Cribari-Neto, F. (2004).  
Beta regression for modelling rates and proportions.  
*Journal of Applied Statistics*, 31(7):799–815.



Ferrari, S. L. P., Espinheira, P. L., and Cribari-Neto, F. (2011).  
Diagnostic tools in beta regression with varying dispersion.  
*Statistica Neerlandica*, pages no–no.



McCullagh, P. and Nelder, J. A. (1989).  
*Generalized Linear Models*.  
Chapman & Hall, London, 2nd edition.

# Bibliography II



Ospina, R., Cribari-Neto, F., and Vasconcellos, K. L. P. (2006).  
Improved point and interval estimation for a beta regression model.  
*Computational Statistics & Data Analysis*, 51(2):960–981.



Ospina, R. and Ferrari, S. (2010).  
Inflated beta distributions.  
*Statistical Papers*, 51:111–126.  
10.1007/s00362-008-0125-4.



Ospina, R. and Ferrari, S. L. P. (2011).  
A general class of zero-or-one inflated beta regression models.  
submitted.



Prater, N. H. (1956).  
Estimate gasoline yields from crudes.  
*Petroleum Refiner*, 35(5):236–238.







Ramsey, J. B. (1969).  
Tests for specification error in classical linear least squares regression analysis.  
*Journal of the Royal Statistical Society B*, 31:350–371.



Rocha, A. V. and Simas, A. B. (2010).  
Influence diagnostics in a general class of beta regression models.  
*Test*.

# Bibliography III

-  Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366.
-  Vasconcellos, K. L. P. and Cribari-Neto, F. (2005). Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*, 19(1):13–31.
-  Wei, B.-C., Hu, Y.-Q., and Fung, W.-K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, 25(1):25–37.
-  Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16.



Thanks !!!!