A Model Based on the Beta Distribution to Deal with Rating Scales

Cedric Taverne - Philippe Lambert

cedric.taverne@uclouvain.be - p.lambert@ulg.ac.be

Institut de Statistique, Biostatistique et sciences Actuarielles

ISBA $\in$ *IMMAQ* $\in$ *UCL*

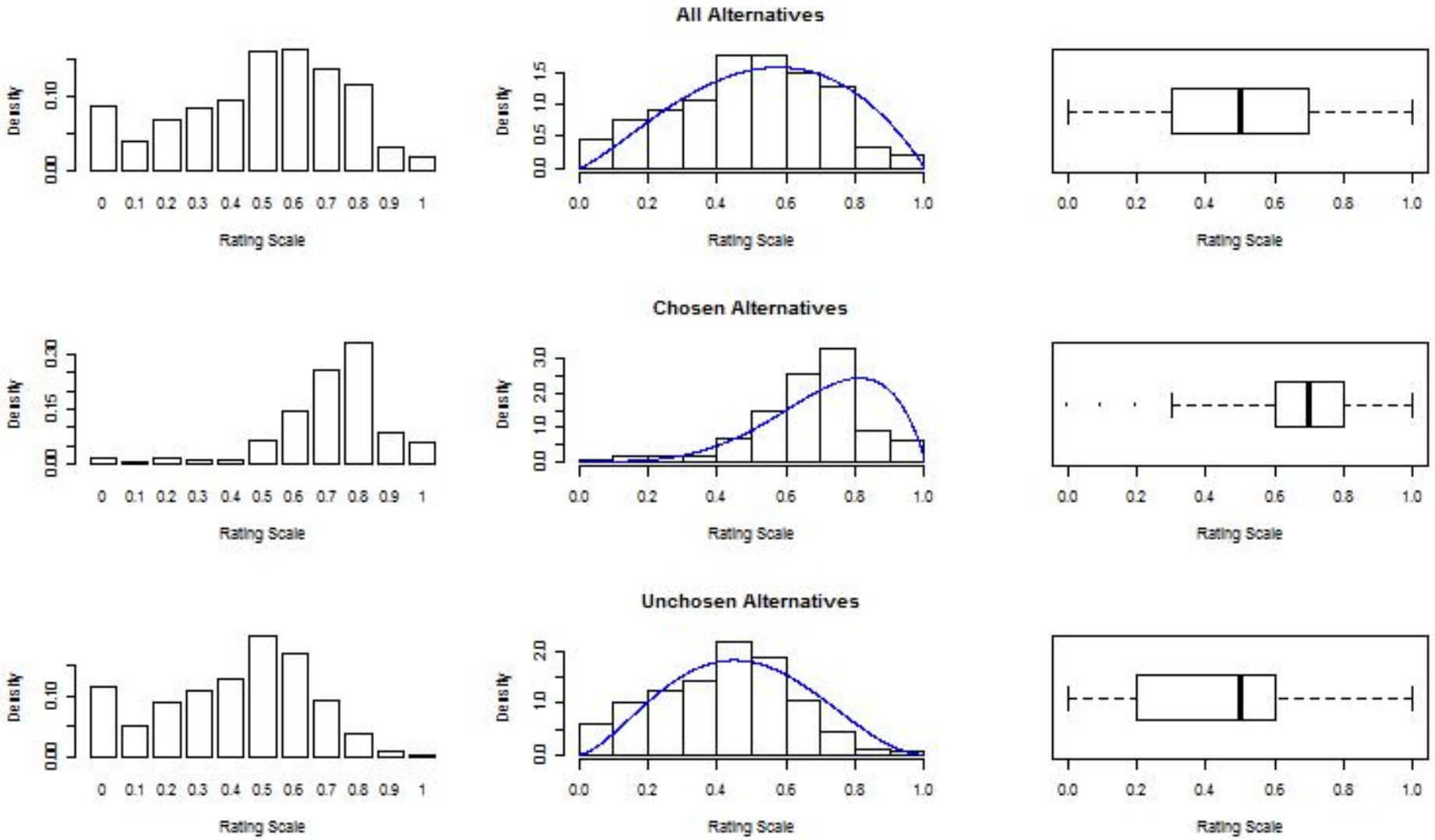
Young Researchers Day - September 23, 2011

Outline

Or Both! (SMCS - Medi-Info)

	Drug 1	**Drug 2**	**Drug 3**	
Medicine :	Original on prescription	Generic on prescription	Original on prescription	
Price per month :	55 € per m.	35 € per m.	75 € per m.	
Regular side effects :	Headaches	Diarrhea	None	
Services :	None	A recipe book	1 hour with a personal coach for free	
Your preference :	☐	☐	☐	☐
Your evaluation :	+-+-+-+-+-+-+-+-+-+ 0 10	+-+-+-+-+-+-+-+-+-+ 0 10	+-+-+-+-+-+-+-+-+-+ 0 10	
	Drug 1	Drug 2	Drug 3	None of these

The Shape of the Rates



What is the Beta Regression?

- Ferrari and Cribari-Neto (2004) have proposed a regression model where the response is beta distributed and the mean response is modeled through a logit link.

$$Y \sim Be(p, q) \qquad p > 0, q > 0,\ y \in (0, 1)$$

$$E(Y) = \frac{p}{p+q} = \mu \qquad Var(Y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1}$$

Let $y = (y_1, ..., y_N)'$ be a random sample, where $y_i \sim \mathcal{B}e(\mu_i, \phi_i)$.

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i'\beta \quad \leftrightarrow \quad \mu_i = \frac{1}{1+e^{-x_i'\beta}} \qquad g : (0,1) \to \mathbb{R}$$

- Ospina *et al.* (2006) have improved point and interval estimation to reduce the bias. Espinheira *et al.* (2008a) have developed influence diagnostics. Espinheira *et al.* (2008b) also discussed the definition of the residuals.
- Simas *et al.* (2010) have added a regression structure on the precision parameter. Since $\phi_i \in (0, \infty)$, a logarithmic link is used.

$$E(Y) = \frac{p}{p+q} = \mu \qquad Var(Y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1}$$

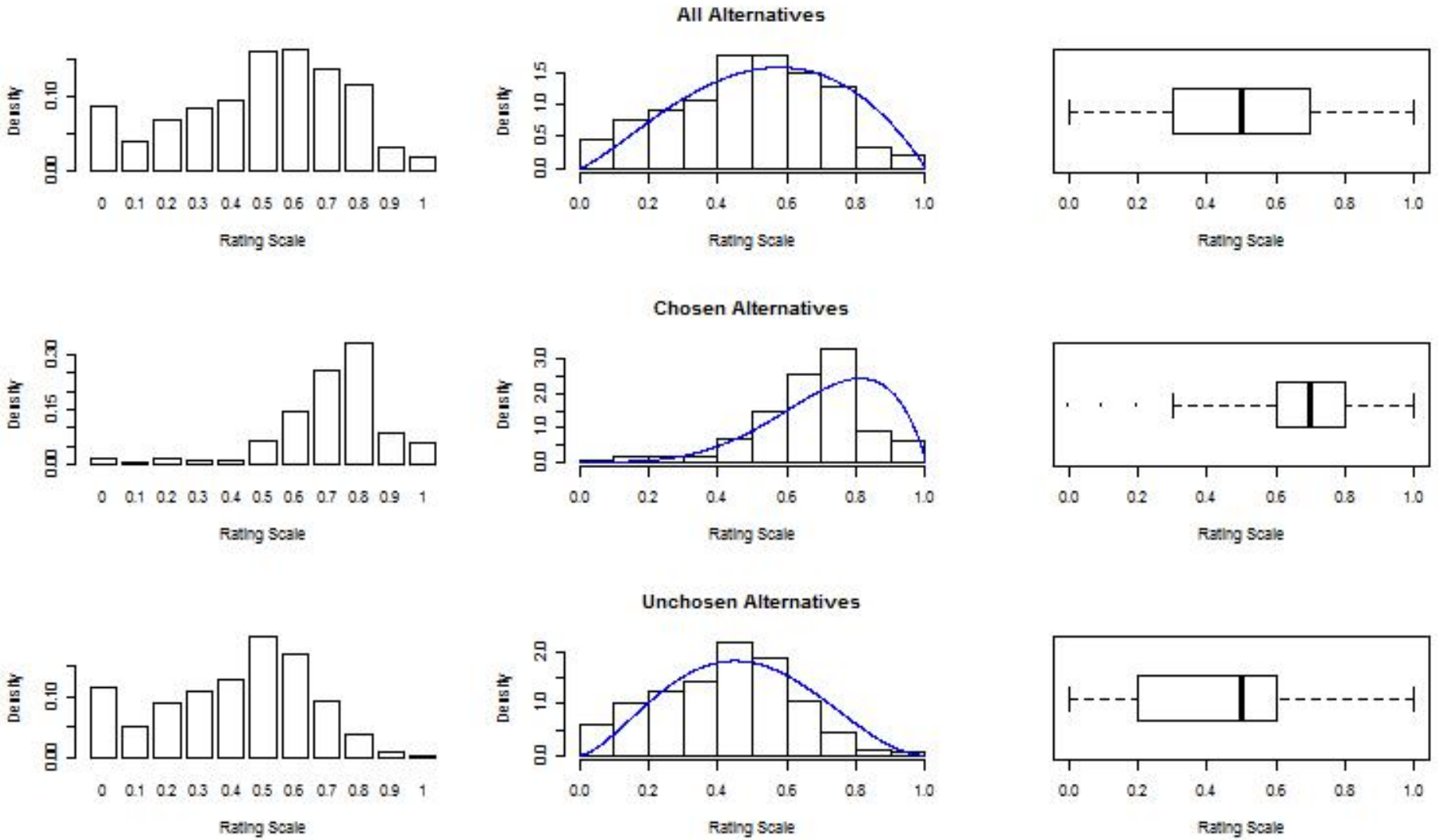
Let $y = (y_1, ..., y_N)'$ be a random sample, where $y_i \sim \mathcal{B}(\mu_i, \phi_i)$.

$$g_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i'\beta \qquad g_1 : (0,1) \to \mathbb{R}$$

$$g_2(\phi_i) = \log(\phi_i) = z_i'\theta \qquad g_2 : (0,\infty) \to \mathbb{R}$$

- Cribari-Neto and Zeileis (2010) implemented all previous reinforcements in the `betareg` package in R.
- Branscum *et al.* (2007) proposed a Bayesian transposition of the beta regression without regression structure on the precision parameter.
- Melo *et al.* (2009) generalized the beta regression for compositional data using the Dirichlet distribution.
- Recent improvements: diagnostic plots for selecting variables (Li-Chu 2011), likelihood inference for small-sample (Ferrari and Pinheiro 2011), partially linear single-index regression structure (Weihua *et al.*, 2012).
- Paolino (2001), Vasconcellos and CribariNeto (2005) and Yuehui *et al.* (2005) amongst others have modeled directly the original parameters of the beta distribution, both through a logarithmic link. This kind of model is dealing with the location and the precision together but are harder to interpret.

Remember the Motivating Data



Dealing with the Left Bound

The inflation in zero seems to correspond to a specific behavior.

So, we model the probability of choosing this lower bound through a simple binary logit model coupled with the further beta regression.

Let $y^\dagger = (y_1^\dagger, ..., y_N^\dagger)'$ be a random sample and $y_i^\dagger \in [a, b]$ be a realization of a random variable $Y^\dagger$.

$$\log\left[\frac{P(Y_i^\dagger = Y_{Low}^\dagger | W)}{1 - P(Y_i^\dagger = Y_{Low}^\dagger | W)}\right] = w_i'\gamma$$

$$\leftrightarrow P(Y_i^\dagger = Y_{Low}^\dagger | W) = \frac{1}{1 + e^{-w_i'\gamma}}$$

Rescaling the scale

In order to fit a beta distribution, our sample $y^\dagger$ has to be rescaled in the $[0, 1]$ interval.

Then, we define:

$$y_i^\star = \frac{y_i^\dagger - a}{b - a}$$

So that $y_i^\dagger \in [a, b]$ and $y_i^\star \in [0, 1]$.

Linking Discrete Scale with a Continuous Distribution

Despite the previous rescaling, the distribution of $Y^\star$ remains discrete while the beta distribution is continuous.

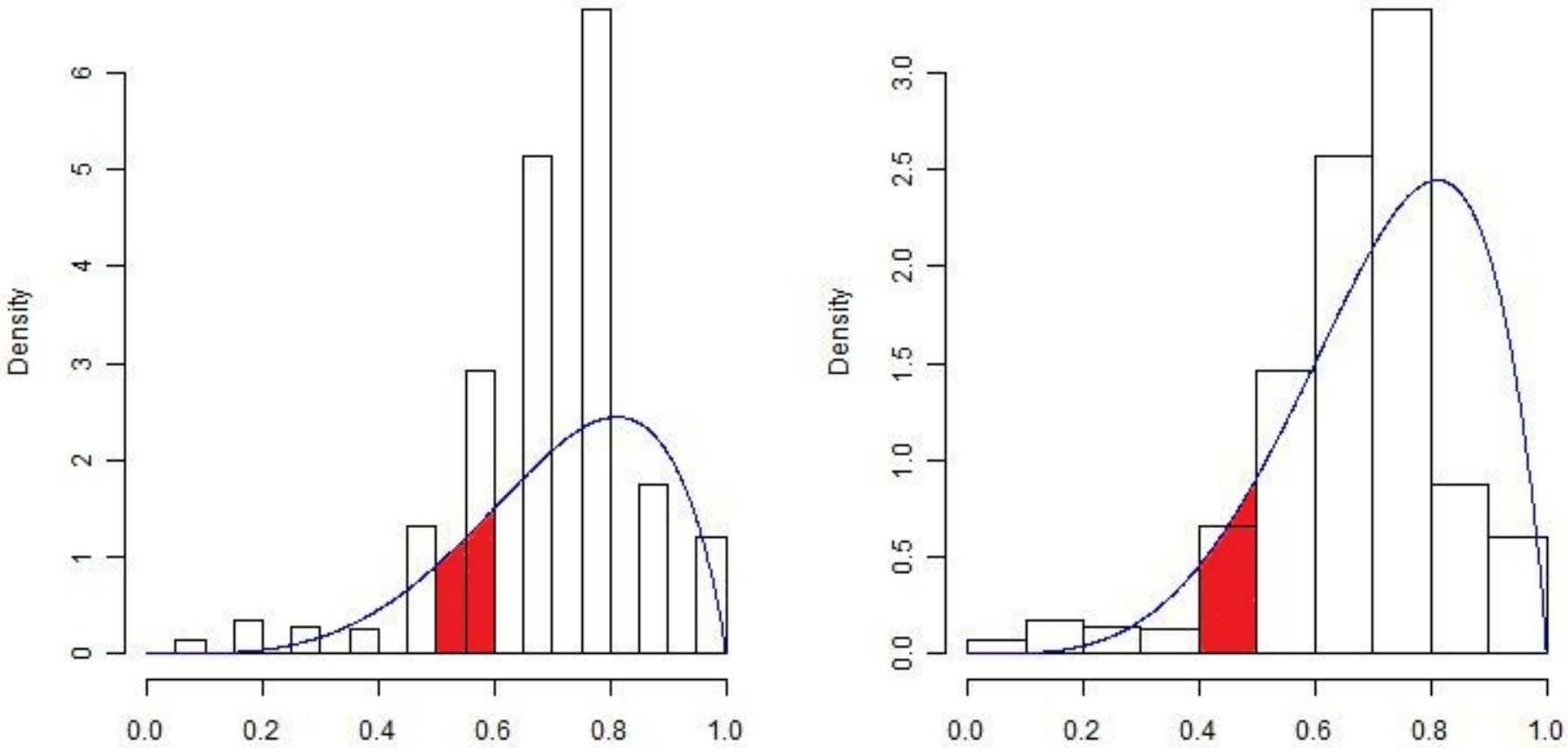
In order to solve that problem, we define k as the distance between each point on the rescaled scale. So that we can define:

$$\begin{aligned} P(Y^\star = y^\star) &= P\left(y_i^\star - k < Y < y_i^\star\right) \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_{y_i^\star - k}^{y_i^\star} u^{p-1}(1-u)^{q-1} du \end{aligned}$$

With $Y \sim Be(p, q)$, $p > 0$, $q > 0$, $y_i^\star \in [k, 1]$ and $k \in (0, 0.5]$.

Linking Discrete Scale with a Continuous Distribution

$$P(Y^\star = y^\star) = P\left(y_i^\star - k < Y < y_i^\star\right)$$



A New Formulation for the Beta Regression

Our formulation of the beta regression is a little bit different:

$$Y \sim Be(p, q) \qquad p, q > 0,\ y \in (0, 1)$$

$$E(Y) = \frac{p}{p+q} = \mu \qquad Var(Y) = \frac{pq}{(p+q)^2} \frac{1}{p+q+1} = \mu(1-\mu)\phi$$

Let $y = (y_1, ..., y_N)'$ be our random sample, where $y_i \sim \mathfrak{B}e(\mu_i, \phi_i)$.

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i'\beta \quad \leftrightarrow \quad \mu_i = \frac{1}{1+e^{-x_i'\beta}} \qquad g_1 : (0,1) \to \mathbb{R}$$

$$g(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = z_i'\theta \quad \leftrightarrow \quad \phi_i = \frac{1}{1+e^{-z_i'\theta}} \qquad g_2 : (0,1) \to \mathbb{R}$$

Consequences of these Changes on the Likelihood

- Contribution of one observation to the likelihood of the beta regression with two regression structures (Simas et al. 2010):

$$\mathcal{L}_i(\mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\,\Gamma((1-\mu_i)\,\phi_i)}\, y_i^{\mu_i\phi_i - 1}\,(1-y_i)^{(1-\mu_i)\phi_i - 1}$$

- Contribution of one observation to the log-likelihood in our model:

$$\mathcal{L}_i(\mu_i, \phi_i) = \frac{\int_{y_i^\star - k}^{y_i^\star} u^{\mu_i\left(\frac{1}{\phi_i}-1\right)-1}\,(1-u)^{(1-\mu_i)\left(\frac{1}{\phi_i}-1\right)-1}\,du}{\mathrm{B}\left(\mu_i\left(\frac{1}{\phi_i}-1\right),(1-\mu_i)\left(\frac{1}{\phi_i}-1\right)\right)}$$

Beta function: $\frac{1}{\mathrm{B}(p,q)} = \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}$

Consequences of these Changes on the Estimation

- Consequences of the definite interval in the likelihood: 8 different constrained expected values appear in the gradient and the Hessian.

  Those values have no closed form and have to be approximated at the individual level for each iteration of the Newton-Raphson algorithm. Consequently, the algorithm is very slow and does not converge easily.

- An alternative approach will be to simulate one point per individual in the interval $\left(y_i^\star - k\ ,\ y_i^\star\right]$ at each iteration. This trick will allow us to drop the integral and go back to the classical beta regression scheme.

Summary

- Deal with the inflation on zero by coupling a binary logit model

$$P(Y_i^{\dagger} = Y_{Low}^{\dagger} | W) = \frac{1}{1 + e^{-w_i' \gamma}}$$

- Adapt the beta regression for discrete rating scales

$$P(Y^{\star} = y^{\star}) \; = \; P\left(y_i^{\star} - k < Y < y_i^{\star}\right) \; , \qquad y_i \sim \mathfrak{Be}(\mu_i, \phi_i)$$

$$\mu_i = \frac{1}{1 + e^{-x_i' \beta}} \qquad \phi_i = \frac{1}{1 + e^{-z_i' \theta}}$$

Next Step

- Finish the work in progress and compare it with other existing models.
- We will develop a Bayesian approach of the beta regression for discrete scales.
- Confirm that the model is adapted for repeated choice task.
- Adapt (sequential) design techniques to the model.

- Branscum, Adam J., and Wesley O. Johnson and Mark C. Thurmond (2007) "Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses", Australian & New Zealand Journal of Statistics, Vol. 49(3), pp.287301.
- Cribari-Neto, Francisco, and Achim Zeileis (2010) "Beta Regression in R", Journal of Statistical Software, Vol. 34(2), pp.124.
- Espinheira, Patricia L., and Silvia L.P. Ferrari and Francisco Cribari-Neto (2008a) "Influence diagnostics in beta regression", Computational Statistics & Data Analysis, Vol. 52(9), pp.4417-4431.
- Espinheira, Patrícia L., and Silvia L. P. Ferrari and Francisco Cribari-Neto (2008b) "On beta regression residuals", Journal of Applied Statistics, Vol. 35(4), pp.407-419
- Ferrari, Silvia, and Francisco Cribari-Neto (2004) "Beta Regression for Modelling Rates and Proportions", Journal of Applied Statistics, Taylor and Francis Journals, Vol.31(7), pp.799-815.
- Ferrari, Silvia L. P., and Eliane C. Pinheiro (2011) "Improved likelihood inference in beta regression", Journal of Statistical Computation & Simulation, Vol. 81(4), pp.431-443.
- Li-Chu, Chien (2011) "Diagnostic plots in beta-regression models", Journal of Applied Statistics, Vol. 38(8), pp.1607-1622.
- Melo, Tatiane F.N., and Klaus L.P. Vasconcellos and Artur J. Lemonte (2009) "Some restriction tests in a new class of regression models for proportions", Computational Statistics & Data Analysis, Vol. 53(12), pp.3972-3979.
- Ospina, Raydonal, and Francisco Cribari-Neto and Klaus L. P. Vasconcellos (2004) "Improved point and interval estimation for a beta regression model", Computational Statistics & Data Analysis, Vol. 51(2), pp.960-981.
- Paolino, Philip (2001) "Maximum likelihood estimation of models with beta-distributed dependent variables", Political Analysis, Vol. 9(4), pp.325-346.
- Simas, Alexandre B., and Wagner Barreto-Souza and Andréa V. Rocha (2010) "Improved estimators for a general class of beta regression models", Computational Statistics & Data Analysis, Vol. 54(2), pp.348-366.
- Vasconcellos, Klaus L.P., and Francisco Cribari-Neto (2005) "Improved maximum likelihood estimation in a new class of beta regression models", Brazilian Journal of Probability and Statistics, Vol. 19, pp.1331.
- Weihua, Zhao, and Zhang Riquan and Huang Zhensheng and Feng Jingyan (2012) "Partially linear single-index beta regression model and score test", Journal of Multivariate Analysis, Vol. 103(1), pp.116-123.
- Yuehui, Wu, and Valerii V. Fedorov and Kathleen J. Propert (2005) "Optimal Design for Dose Response Using Beta Distributed Responses", Journal of Biopharmaceutical Statistics, Vol. 15(5), pp.753-771