MRC Chalk Talk

Proportions and Rates as Dependent Variables

Yoonsang Kim, PhD

Methodology Research Core
Institute for Health Research and Policy
University of Illinois at Chicago

Jan 14, 2014

- Proportions or rates as dependent variables
  - Fraction of total weekly hours spent working
  - Proportion of heart transplant performed at hospitals last year
  - Proportion of income spent on food
- Motivation
- Beta Regression
- Fractional Logit Model
- Software

Beta Regression

- Dependent variables bounded in [0,1]
- Skewness
- Heteroskedasticity
- Fit linear regression assuming normality?
  - Inaccurate interval estimation/hypothesis testing
- Transform the dependent variable?
  - Log or logit transformation
  - Model the mean of transformed response
  - $\mathrm{E}\left(\log \frac{y}{1-y}\right) \neq \log \frac{E(y)}{1-E(y)}$

Beta Regression

- Ferrari & Cribari-Neto (2004)
- Beta distribution:

$$f(y;\omega,\tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)+\Gamma(\tau)}\, y^{\omega-1}(1-y)^{\tau-1}$$

where $0 < y < 1;\ \omega > 0\ ; \tau > 0$

$\Gamma(t) = \int x^{t-1}e^{-x}dx$; Gamma function

$$E(Y) = \frac{\omega}{\omega+\tau}$$

$$Var(Y) = \frac{\omega\tau}{(\omega+\tau)^2(\omega+\tau+1)}$$

two-parameter exponential family

Beta Regression

- Re-parameterization:

$$E(Y) = \frac{\omega}{\omega + \tau} = \mu$$

$$Var(Y) = \frac{\omega\tau}{(\omega + \tau)^2(\omega + \tau + 1)} = \frac{\mu(1-\mu)}{1+\phi}$$

Beta Regression

- Re-parameterization:

$$E(Y) = \frac{\omega}{\omega + \tau} = \mu$$

$$Var(Y) = \frac{\omega\tau}{(\omega + \tau)^2(\omega + \tau + 1)} = \frac{\mu(1-\mu)}{1+\phi}$$

$$f(y; \mu, \phi)$$

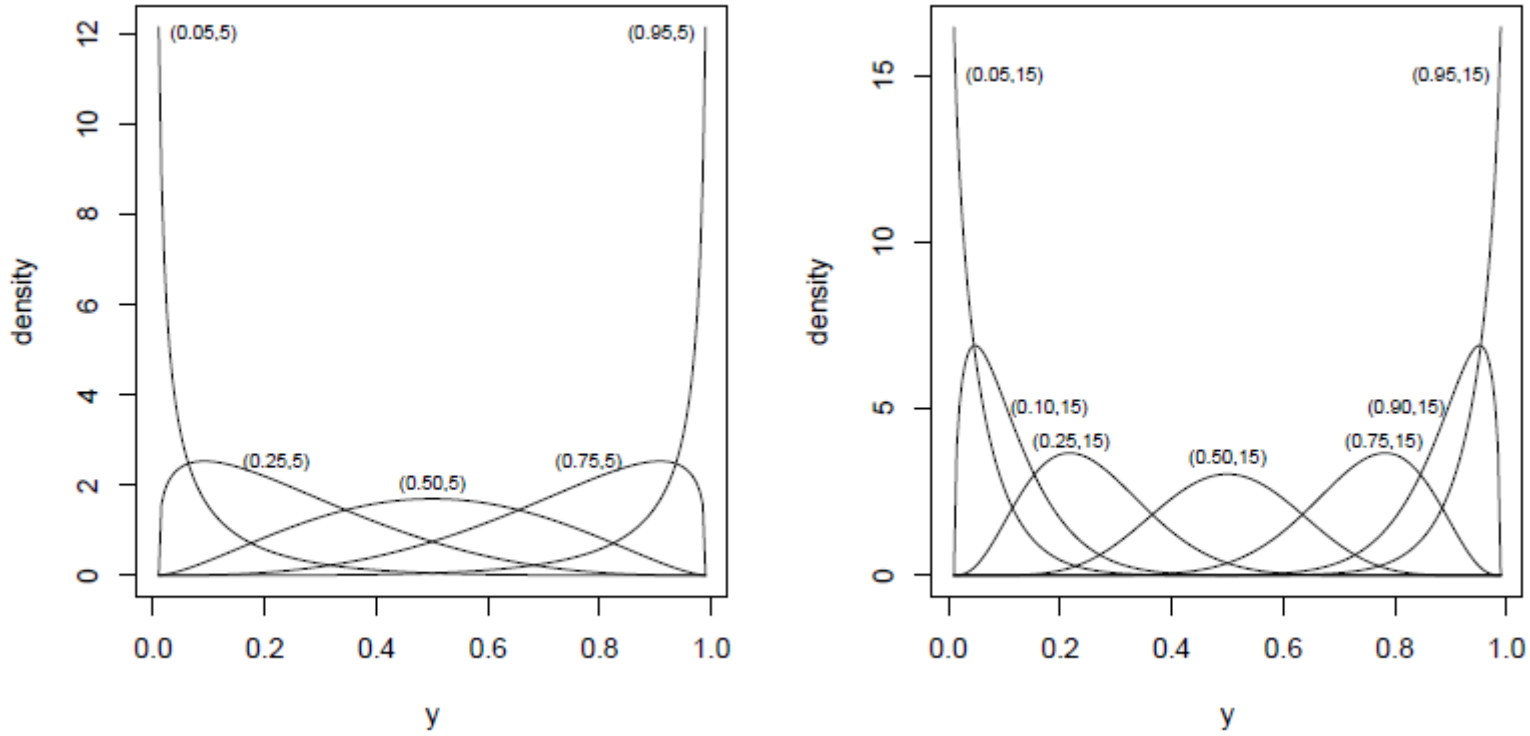
$$= \frac{\Gamma(\phi)}{\Gamma(\mu\phi) + \Gamma((1-\mu)\phi)} \, y^{\mu\phi - 1}(1-y)^{(1-\mu)\phi - 1}$$

where $0 < \mu < 1; \phi > 0$

Beta Regression

Beta distributions with different parameter values

$$B(\mu, \phi)$$



Beta Regression

- The model:

$$\log\frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - Link function = "logit" maps $(0,1)$ to $(-\infty, \infty)$
  - For $0 < y < 1$
  - Non-constant variance b/c $Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$
  - $\phi =$ precision parameter
  - Similar to GLM framework (link function, linear predictor, dispersion)
  - Possible link functions: probit, log-log, clog-log

Beta Regression

- Estimation:
  - Maximum likelihood estimation for (μ, ϕ)
  - LR test, Score test, Wald test (Ferrari & Cribari-Neto, 2004)
- Diagnostic Measures:
  - Pseudo R^2
  - Residuals: deviance, Pearson, standardized weighted residual 2 (Espinheira, Ferrari, & Cribari-Neto, 2008)
- Software:
  - R betareg
  - Stata betafit
  - SAS Glimmix specifying dist=beta

Beta Regression

- Limitation:

  y doesn't include 0 and 1

  Transform $y^* = \frac{1}{n}(y(n-1) + 0.5)$

- Variable dispersion beta regression:

  Additional to the location model, the dispersion model is

  $$\log \phi = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_q z_q$$

  Smithson & Verkuilen (2006) used $-\gamma_i$

- Different link functions can improve the model-fit

Beta Regression

Example:

Household food expenditures for 38 households (Griffiths et al. 1993).

$y =$ the proportion of household income spent on food

$x1 =$ # persons

$x2 =$ income

```
betareg(formula = I(food/income) ~ income + persons, data = FoodExpenditure, link
= "logit")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.7818 -0.4445  0.2024  0.6852  1.8755

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.622548   0.223854  -2.781 0.005418 **
income      -0.012299   0.003036  -4.052 5.09e-05 ***
persons      0.118462   0.035341   3.352 0.000802 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)    35.61       8.08   4.407 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 45.33 on 4 Df
Pseudo R-squared: 0.3878
Number of iterations: 28 (BFGS) + 4 (Fisher scoring)
```

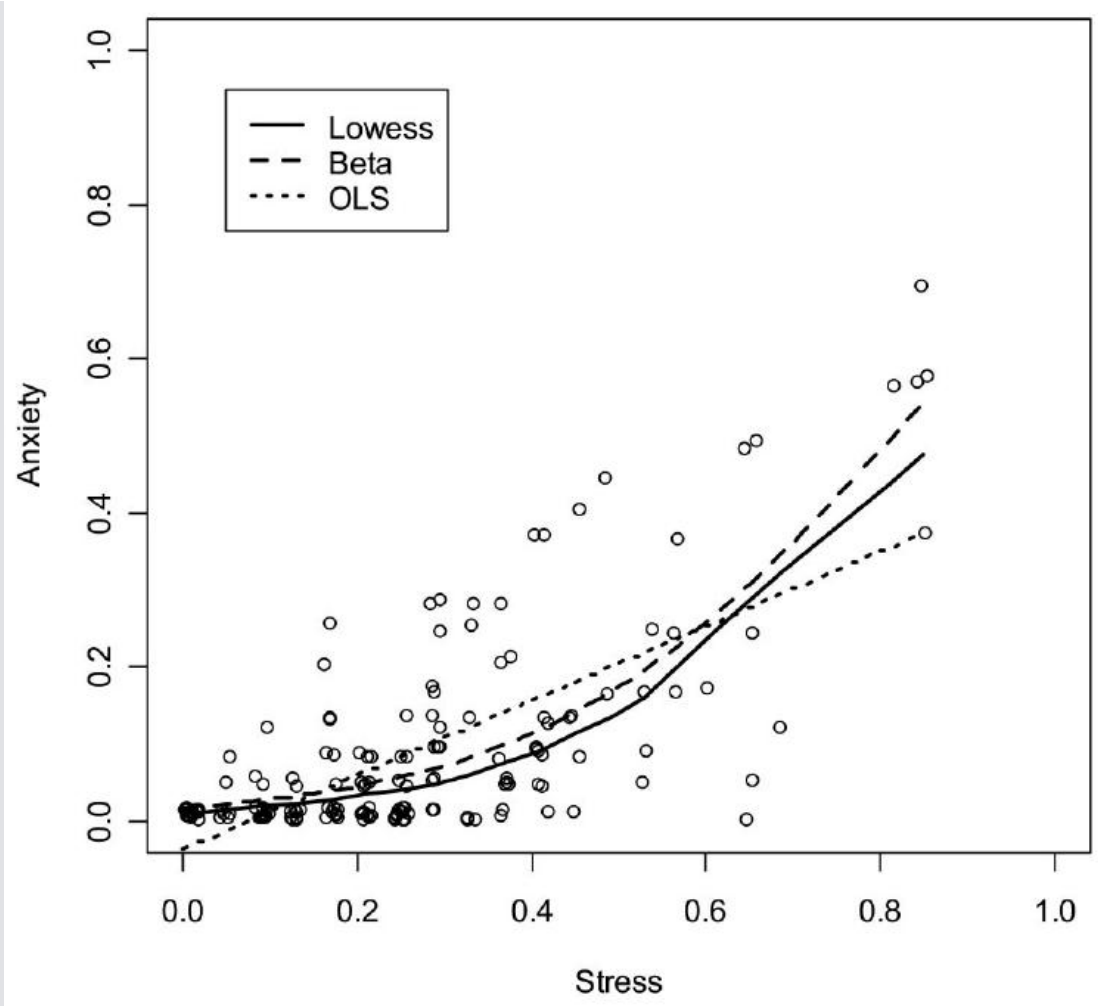
Beta Regression

Variable dispersion beta regression

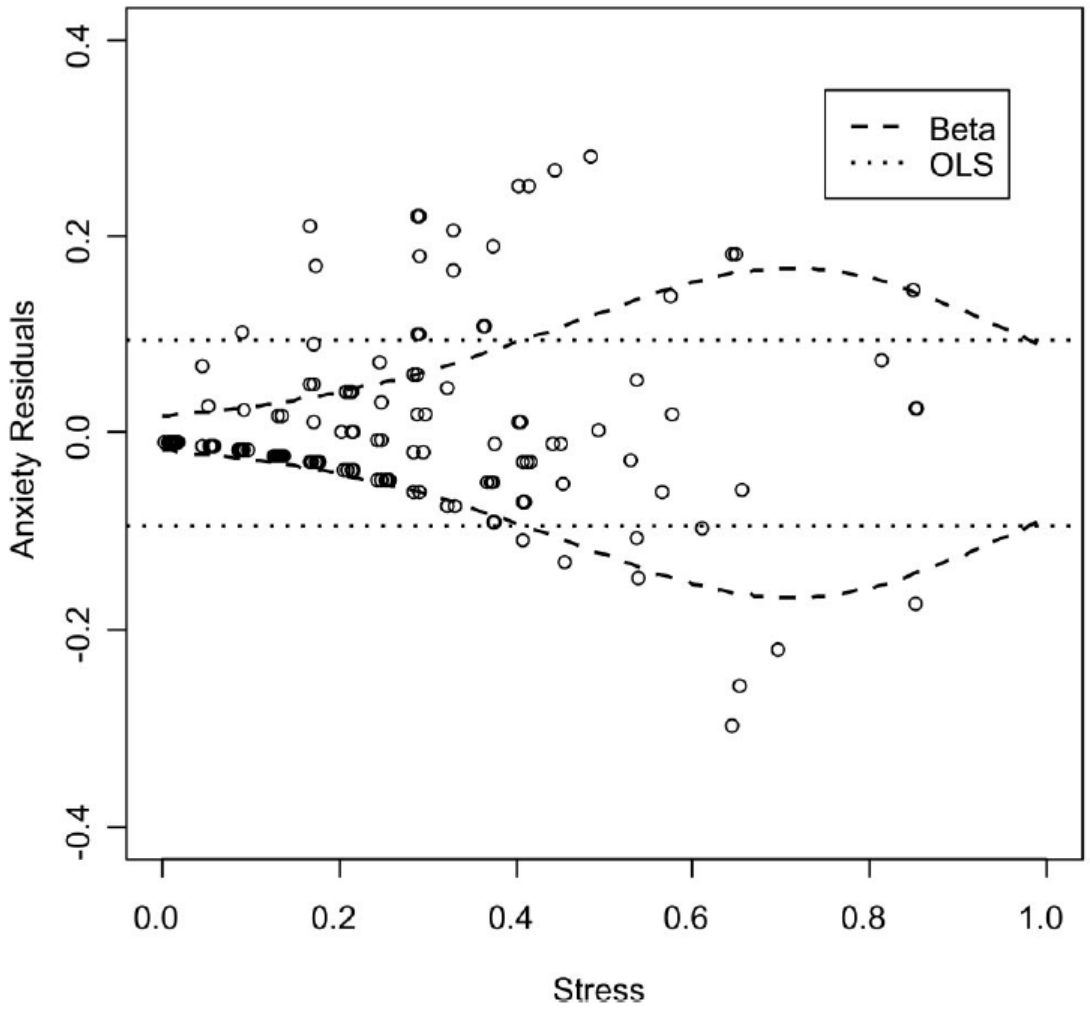
```
betareg(formula = I(food/income) ~ income + persons | persons, data =
FoodExpenditure, link = "logit")

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.8660 -0.7478  0.3549  0.7825  1.8947

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.783082   0.177708  -4.407 1.05e-05 ***
income      -0.008217   0.002411  -3.409 0.000653 ***
persons      0.092554   0.034821   2.658 0.007862 **

Phi coefficients (precision model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.5043     0.5334  10.320  < 2e-16 ***
persons      -0.4835     0.1335  -3.623 0.000291 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 49.18 on 5 Df
Pseudo R-squared: 0.3852
Number of iterations: 20 (BFGS) + 1 (Fisher scoring)
```

Beta Regression

Stress and Anxiety (Smithson & Verkuilen, 2006)



Beta Regression



Beta Regression

Other extended models:

- Bias-corrections for fixed dispersion beta regressions (Ospina, Cribari-Neto, & Vasconcellos, 2006)
- Bayesian approach to beta regression (Branscum, Johnson, & Thurmond, 2007)
- Zero-inflated beta regression by incorporating degenerate distributions to model the extreme values (Cook, Kieschnick, & McCullough, 2008;Ospina & Ferrari, 2012)
- Truncated inflated beta regression: allow truncation in a subset [c,1] of the unit interval and mass points at c, zero, and one (Pereira, Botter, & Sandoval, 2012)
- Random effect beta regression (Bonat et al. Not published)

Fractional Logit Model

- The Model:

  Distribution of y?

  GLM framework as if the response $\sim$ binomial distr

  (Papke & Wooldridge, 1996)

  Log-likelihood:

$$\ell = y \log G(X\beta) + (1 - y) \log[1 - G(X\beta)]$$

  Link function $0 < G(\cdot) < 1$

  $G(X\beta) = \mu$

  Works for $0 \leq y \leq 1$

  With logit link

$$\log \frac{\mu}{1 - \mu} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Fractional Logit Model

- Estimation

  Maximum quasi-likelihood estimation (McCullagh & Nelder, 1998)

- Variance

  GLM framework: $Var(y) = \sigma^2 \mu(1-\mu)$

  The constant over-dispersion parameter (σ^2) may "fail" when X is not independent of group sizes

  e.g.X = firm characteristic

  Y = #workers contributing to 401(k)/#workers at firm,

  → Robust "sandwich" estimator for β

Fractional Logit Model

- Software
  - SAS Glimmix with dist=binomial
  - Stata glm command with family(binomial), link(logit), vce(robust)
  - Stata fractional logit module
  - R function glm, apply the sandwich function

References

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 & New Zealand Journal of Statistics*, *49*(3), 287–301. doi:10.1111/j.1467-842X.2007.00481.x

Cook, D. O., Kieschnick, R., & McCullough, B. D. (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, *15*(5), 860–867. doi:10.1016/j.jempfin.2008.02.001

References

Espinheira, P. L., Ferrari, S. L. P., & Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, *35*(4), 407–419. doi:10.1080/02664760701834931

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

McCullagh, P., & Nelder, J. A. (1998). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

References

Ospina, R., Cribari-Neto, F., & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis*, *51*(2), 960–981. doi:10.1016/j.csda.2005.10.002

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. doi:10.1016/j.csda.2011.10.005

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, *11*(6), 619–632. doi:10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

References

Pereira, G. H. A., Botter, D. A., & Sandoval, M. C. (2012). The Truncated Inflated Beta Distribution. *Communications in Statistics - Theory and Methods*, *41*(5), 907–919. doi:10.1080/03610926.2010.530370

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. doi:10.1037/1082-989X.11.1.54