`PLreg`: an R package for modeling bounded continuous data

Francisco F. Queiroz, Silvia L. P. Ferrari
University of São Paulo

Abstract: Bounded continuous data, particularly on the unit interval, appear in various research areas, including medicine, biology, sociology, psychology, economics, among many others. This work introduces the new R `PLreg` package which allows fitting the power logit regression models for bounded continuous data, in which the density generator may be normal, Student-t, power exponential, slash, hyperbolic, sinh-normal, or type II logistic. Some particular cases and extensions of the power logit model are also implemented in the package. Diagnostic tools associated with the fitted model, such as the residuals, local influence measures, leverage measures, and goodness-of-fit statistics, are available. The estimation process follows the maximum likelihood approach. The package is applied to real data sets for illustration.

Related publications:

– Queiroz and Ferrari, Statistical Modelling, to appear (2022)
– Lemonte and Bazán, Biometrical Journal, 58, 727–746 (2016)

Introduction: power logit regression models

- **Context**: predict or explain the behavior of a continuous proportion from a set of other variables.

Power logit distribution

$$Z = \frac{1}{\sigma}\left[\text{logit}\left(Y^{\lambda}\right) - \text{logit}(\mu^{\lambda})\right] \sim \text{S}(0, 1; r)$$

Y has a power logit (PL) distribution with parameters $0 < \mu < 1$, $\sigma > 0$, and $\lambda > 0$, representing the median, dispersion, and skewness of Y, respectively.

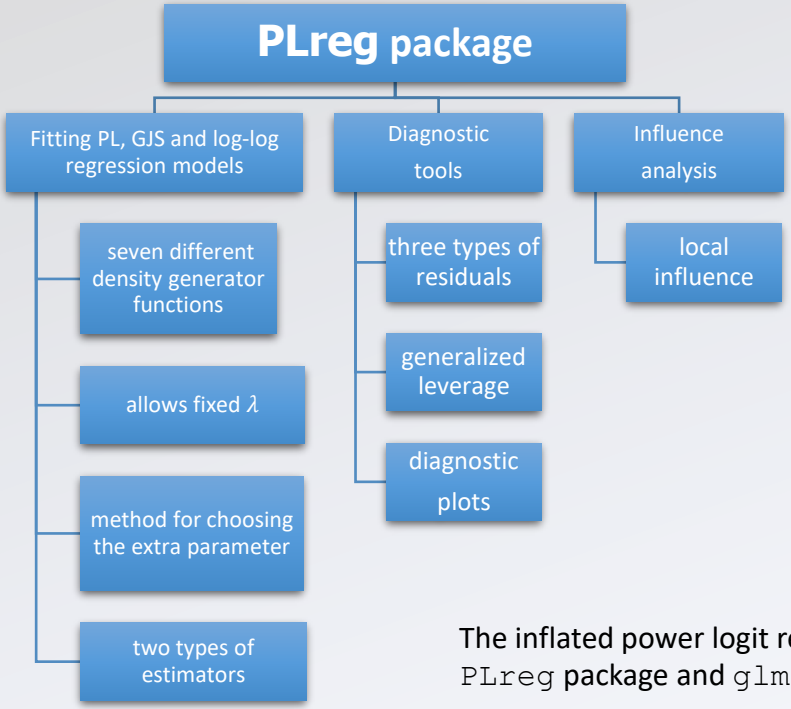
Definition and estimation

- $Y_1, \ldots, Y_n$ independent random variables with $Y_i \sim \text{PL}(\mu_i, \sigma_i, \lambda; r)$.
- $r(\cdot)$ is the density generator function which may depend on an extra parameter.
- $d_1(\mu_i) = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}$, $d_2(\sigma_i) = \boldsymbol{s}_i^{\top}\boldsymbol{\tau}$ where $d_1(\cdot)$ and $d_2(\cdot)$ are the link functions, $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$ are the vectors of covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are the parameter vectors.
- GJS regression models: $\lambda = 1$; log-log regression models ($\lambda \to 0$).
- The estimation process is based on maximum likelihood approach.

- Flexible;
- Direct parameter interpretation;
- Extensions: inflated power logit regression models: [0,1) or (0,1]; see Queiroz (2022).

PLreg package on CRAN repository in R

`PLreg` package



Main functions

- `PLreg(formula, data, subset, na.action, Family, zeta = NULL, link, link.sigma = NULL, type = c("pML", "ML"), control = PLreg.control(...), ...)`
- `dPL(x, ...)`, `pPL(q, ...)`, `qPL(p, ...)`, and `rPL(n, ...)`
- `extra.parameter(object, lower, upper, grid)`
- **Some available methods:** `plot()`, `summary()`, `AIC()`, `BIC()`, `vocv()`.

The inflated power logit regression model can also be fitted by using jointly `PLreg` package and `glm()` function.

Illustrative example

```
> fitPL <- PLreg(firmcost ~ sizelog + indcost,
                 data = Firm,
                 family = "SLASH",
                 zeta = 2.33)
> summary(fitPL)
```

```
> extra.parameter(fitPL, 1, 4)

> plot(fitPL, which = 7)
```

```
> summary(fitPL)

Call:
PLreg(formula = firmcost ~ sizelog + indcost, data = Firm, family = "SLASH", zeta = 2.33)

Standardized residuals:
    Min      1Q  Median      3Q     Max
-2.1196 -0.6613  0.0548  0.7173  5.9008

Coefficients (median model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.8625     0.9994   3.865 0.000111 ***
sizelog      -0.9049     0.1120  -8.077 6.66e-16 ***
indcost       2.1345     0.5835   3.658 0.000254 ***

Sigma coefficients (dispersion model with log link):
       Estimate Std. Error z value Pr(>|z|)
(sigma)   0.1481     0.5344   0.277    0.782

Lambda coefficient:
         Estimate Std. Error
(lambda)    1.809      1.022
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Family: PL - SLASH ( 2.33 ) (Power logit slash)
Estimation method: pML (penalized maximum likelihood)
Log-likelihood: 123.7 on 5 Df
Pseudo R-squared: 0.4162
Upsilon statistic: 0.06446
AIC: -237.4
Number of iterations in BFGS optimization: 15
```