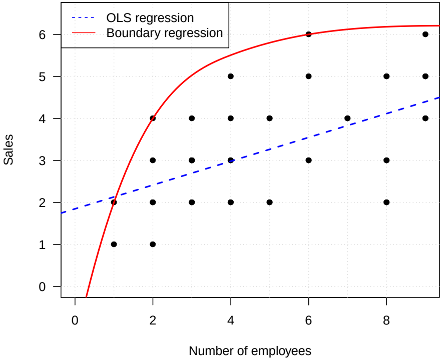
npbr: A Package for Nonparametric Boundary Regression in R

Abdelaati Daouia[1] Thibault Laurent[1,2]
Hohsuk Noh[3]

[1]TSE: Toulouse School of Economics, University of Toulouse Capitole, France
[2]CNRS [3]Sookmyung Women's University

useR!2018
13^{th} July 2018, Brisbane, Australia

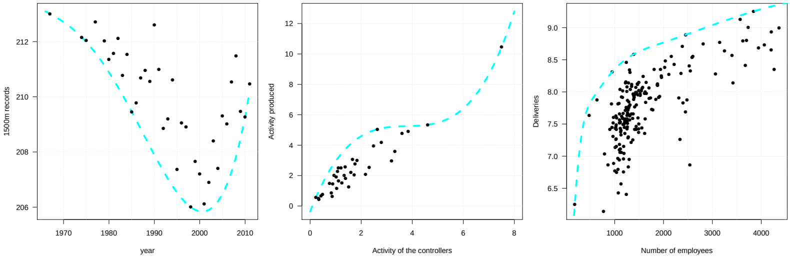
What does "boundary regression" mean ?



Theory of production (economics)

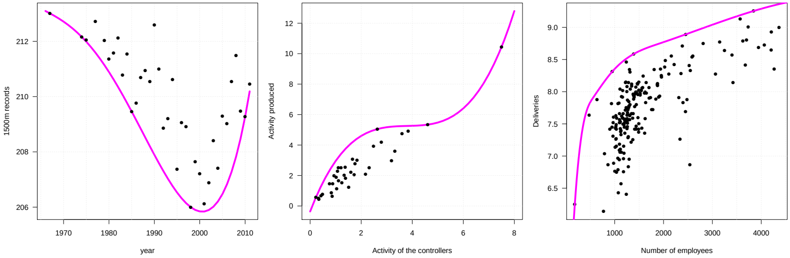
- X: input
- Y: output
- n observations $(x_1, y_1), \ldots, (x_n, y_n)$
- Def: maximum producible quantity of Y for any given quantity of X
- φ: boundary (or frontier)

Example of data



- Annual sport records: `data("records")`
- European air controllers: `data("air")`
- French postal services: `data("post")`

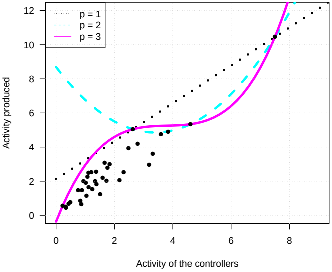
Constrained boundary regression



- Unconstrained: `method = "u"`
- Monotone: `method = "m"`
- Monotone and concave: `method = "mc"`

Parametric model: polynomial estimators

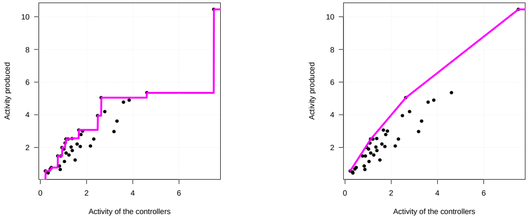
- Model structure: $\varphi_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_p x^p$
- Optimization problem: Find $\hat{\theta} = (\hat{\theta}_0, \ldots, \hat{\theta}_p)$ s.t. that $\varphi_{\hat{\theta}}$ envelopes the full data and minimizes the area under its graph (Hall, 1998)



- Choose p which minimizes the AIC/BIC (Daouia, 2015)
- Easily implemented, but no constained version

Nonparametric model: FDH, DEA

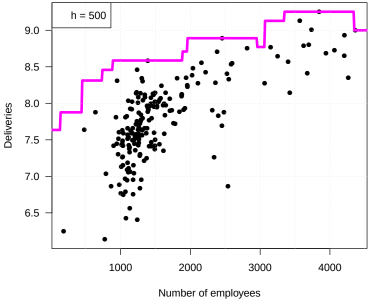
- Nonparametric: model structure is not specified *a priori* but is instead determined from data
- Example: Free Disposal Hull (Deprins, Simar and Tulkens, 1984) or Data Envelopment Analysis (Farrell, 1957)



- Very famous and popular in the economic litterature, but too sensitive to the extreme values

Nonparametric model: Local maximum frontier estimators

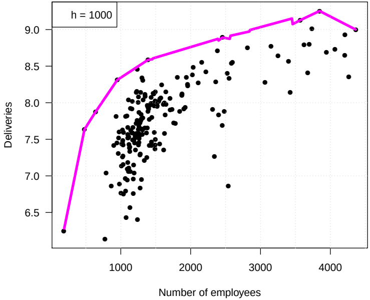
- $\tilde{\varphi}(x) = \max_{i=1,\ldots,n} y_i \mathbf{1}_{\{|x_i - x| \le h\}}$ (Gijbels and Peng, 2000)
- A data-driven rule for selecting h: package **np** (Li, Lin and Racine, 2013)



- No constrained version and h should depend on x

Nonparametric model: Localized linear fitting

- $\hat{\varphi}_{n,LL}(x) = \min\{z : \text{there exists } \theta \text{ s.t. } y_i \leq z + \theta(x_i - x) \text{ for all } i \text{ s.t. } x_i \in (x - h, x + h)\}$, (Hall *et al.*, 1998)
- Optimal h: Hall and Park (2004)



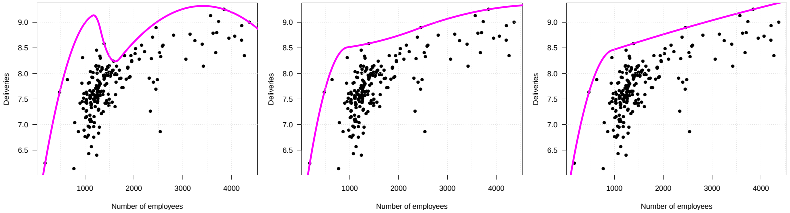
- No constained version

Nonparametric model: Quadratic (or Cubic) spline fitting

Daouia *et al.* (2016)

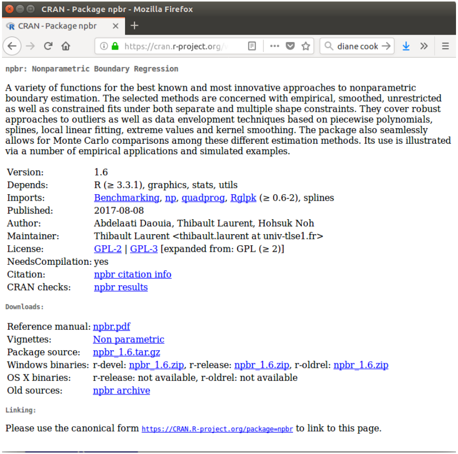
- Denote a partition of $[a,b]$ by $a = t_0 < t_1 < \ldots < t_{k_n} = b$, with $a = \min_i x_i$ and $b = \max_i x_i$ by considering the j/k_nth quantiles $t_j = x_{[jn/k_n]}$ of the distinct values of x_i for $j = 1, \ldots, k_n - 1$
- Let $N = k_n + 2$ and $\pi(x) = (\pi_0(x), \ldots, \pi_{N-1}(x))^\top$ be the vector of normalized B-splines of order 2 (or 3) based on $\{t_j\}$
- $\hat{\varphi}_n(x) = \pi(x)^\top \hat{\alpha}$, where $\hat{\alpha}$ minimizes the same objective function as $\tilde{\alpha}$ subject to the same envelopment constraints and the additional monotonicity constraints $\pi'(t_j)^\top \alpha \geq 0$, $j = 0, 1, \ldots, k_n$, with π' being the derivative of π.
- Number of inter-knot segments k_n: AIC or BIC

Quadratic spline fitting



- Unconstrained ($k_n = 3$)
- Monotone ($k_n = 2$)
- Monotone and concave ($k_n = 1$)

Package on CRAN



CRAN - Package npbr - Mozilla Firefox

CRAN - Package npbr

https://cran.r-project.org/

npbr: Nonparametric Boundary Regression

A variety of functions for the best known and most innovative approaches to nonparametric boundary estimation. The selected methods are concerned with empirical, smoothed, unrestricted as well as constrained fits under both separate and multiple shape constraints. They cover robust approaches to outliers as well as data envelopment techniques based on piecewise polynomials, splines, local linear fitting, extreme values and kernel smoothing. The package also seamlessly allows for Monte Carlo comparisons among these different estimation methods. Its use is illustrated via a number of empirical applications and simulated examples.

Version:	1.6
Depends:	R (≥ 3.3.1), graphics, stats, utils
Imports:	Benchmarking, np, quadprog, Rglpk (≥ 0.6-2), splines
Published:	2017-08-08
Author:	Abdelaati Daouia, Thibault Laurent, Hohsuk Noh
Maintainer:	Thibault Laurent <thibault.laurent at univ-tlse1.fr>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	yes
Citation:	npbr citation info
CRAN checks:	npbr results

Downloads:

Reference manual:	npbr.pdf
Vignettes:	Non parametric
Package source:	npbr_1.6.tar.gz
Windows binaries:	r-devel: npbr_1.6.zip, r-release: npbr_1.6.zip, r-oldrel: npbr_1.6.zip
OS X binaries:	r-release: not available, r-oldrel: not available
Old sources:	npbr archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=npbr to link to this page.

On MAC OS: install the glpk library outside of R, using homebrew

npbr functions

Function	Description	Reference
dea_est	DEA, FDH and linearized FDH	Farrell (1957), Deprins *et al.* (1984), Hall and Park (2002)
loc_est	Local linear fitting	Hall *et al.* (1998), Hall and Park (2004)
loc_est_bw	Bandwidth choice for local linear fitting	Hall and Park (2004)
poly_est	Polynomial estimation	Hall (1998)
poly_degree	Optimal polynomial degree selection	Daouia *et al.* (2015)
dfs_momt	Moment type estimation	Daouia *et al.* (2010), Dekkers *et al.* (1989)
dfs_pick	Pickands type estimation	Daouia *et al.* (2010), Dekkers *et al.* (1989)
rho_momt_pick	Conditional tail index estimation	Daouia *et al.* (2010), Dekkers *et al.* (1989)
kopt_momt_pick	Threshold selection for moment/Pickands frontiers	Daouia *et al.* (2010)
dfs_pwm_regul	Nonparametric frontier regularization	Daouia *et al.* (2012)
loc_max	Local constant estimation	Gijbels and Peng (2000)
pick_est	Local extreme-value estimation	Gijbels and Peng (2000)
quad_spline_est	Quadratic spline fitting	Daouia *et al.* (2015)
quad_spline_kn	Knot selection for quadratic spline fitting	Daouia *et al.* (2015)
cub_spline_est	Cubic spline fitting	Daouia *et al.* (2015)
cub_spline_kn	Knot selection for cubic spline fitting	Daouia *et al.* (2015)
kern_smooth	Nonparametric kernel boundary regression	Parmeter and Racine (2013), Noh (2014)
kern_smooth_bw	Bandwidth choice for	Parmeter and Racine (2013),

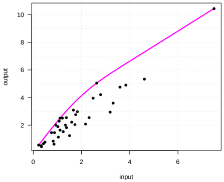
Usage

```
poly_est(xtab, ytab, x, deg, control =
list("tm_limit" = 700))
```

- `xtab`: a numeric vector containing the observed inputs $x_1, \ldots, x_n$.
- `ytab`: a numeric vector of the same length as xtab containing the observed outputs $y_1, \ldots, y_n$.
- `x`: a numeric vector of evaluation points in which the estimator is to be computed.
- `deg`: an integer (polynomial degree).
- `control`: a list of parameters to the GLPK solver.

Example of use

```
R> require("npbr")
R> data(air)
R> x_air <- seq(min(air$xtab), max(air$xtab),
+  length.out = 101)
R> kn <- cub_spline_kn(air$xtab, air$ytab,
+   method = "mc", type = "BIC")
R> y_est <- cub_spline_est(air$xtab, air$ytab,
+    x_air, kn = kn, method = "mc")
R> plot(x_air, y_cub_air_mc, lwd = 3, type = "l",
+    xlab = "input", ylab = "output")
R> points(air$xtab, air$ytab, pch = 16)
```



Conclusion

- Available on CRAN
- Article published in Journal of Statistical Software (2017), `http://dx.doi.org/10.18637/jss.v079.i09`
- Numerical illustrations given in the article: cubic spline fitting seems to be the best method whatever the shape of the data
- Hope to see you in UseR!2019, Toulouse, France, next year