What to expect – an R vignette for expectreg

Fabian Sobotka, Thomas Kneib, Sabine Schnabel, Paul Eilers

October 30, 2024

Abstract

expectreg is an R package for estimating expectile curves from univariate and multivariate data. Expectile curves are a valuable least squares alternative to quantile regression which is based on linear programming techniques. expectreg provides a number of functions for different approaches taken to estimate expectiles investigated since their introduction in [Newey and Powell(1987)] using asymmetric least squares.

1 Overview

This section offers an overview over the functions implemented in expectreg. It assumes that the user already installed the package successfully.

```
> library(expectreg)
```

```
> help(package = "expectreg")
> data(package = "expectreg")
```

will give you a short overview about the available help files of the package as well as the data that will be provided with expectreg. The package includes the following functions:

rb	Creates bases for a regression based on covariates
demq	Density of a special distribution developed by Roger Koenker [Koenker(1992)]
ebeta	Expectiles of the beta distribution
eemq	Expectiles of a special distribution developed by Roger Koenker
enorm	Expectiles of the normal distribution
eunif	Expectiles of the uniform distribution
expectreg.boost	Expectile regression using boosting
expectreg.ls	Expectiles regression of additive models
expectreg.qp	Expectile sheets with monotonicity constraints
pemq	Distribution function for a special distribution developed by Roger Koenker
qemq	Quantile function for a special distribution developed by Roger Koenker
quant.boost	Quantile regression using boosting
remq	Random variable generated from a special distribution developed by Roger Koenker

2 Expectiles in a nutshell

2.1 Introduction to expectiles using LAWS

Asymmetric least squares or least asymmetrically weighted squares (LAWS) is a weighted generalization of ordinary least squares (OLS) estimation. LAWS minimizes

$$S = \sum_{i=1}^{n} w_i(p)(y_i - \mu_i(p))^2,$$

with

$$w_i(p) = \begin{cases} p & \text{if } y_i > \mu_i(p) \\ 1-p & \text{if } y_i \leq \mu_i(p) \end{cases}, \quad (1)$$

where y_i is the response and $\mu_i(p)$ is the population expectile for different values of an asymmetry parameter p with $0 < p < 1$. The model is fitted by alternating between weighted regression and recomputing weights until convergence (when the weights do not change anymore). Equal weights ($p = 0.5$) give a convenient starting point.

For the expectile curve $\mu(p)$ several choices for the functional form are possible. The original proposal in [Newey and Powell(1987)] favored a linear model. We suggest a more flexible functional form for the expectile curve. [Schnabel and Eilers(2009)] proposed to model expectile curves with $P-$splines. Other types such as other splines, markov random field or other options are also possible (see [Sobotka and Kneib(2010)]).

2.2 Expectile bundle model

In theory it is not possible that expectile curves cross, but in estimation practise it is often encountered due to sampling variation. The expectile bundle model is a location-scale type of model that allows for the simultaneous estimation of a set of expectiles. By its construction crossing over of curves is not possible.

In the expectile bundle model the expectiles $\mu(x, p)$ are defined by

$$\mu(x, p) = t(x) + c(p)s(x) \quad (2)$$

where $t(x)$ is a common smooth trend of all expectile curves specified by a P-spline. $c(p)$ is the asymmetry function of the bundle describing the spread, i.e. the set of standardized expectiles. $s(x)$ represents the local width of the expectile bundle and is also formulated as a P-spline. The estimation procedure consists of two steps. In Step 1 the common trend $t(x)$ is estimated. Then in step 2 we use the detrended response $y - t(x)$ to estimate $s(x)$ and $c(p)$ in an iterative procedure.

The expectiles bundle model is explained in more detail in [Schnabel and Eilers(2010)].

2.3 Restricted regression quantiles

In [He(1997)] proposed a version of restricted regression quantiles to avoid the crossing of quantile curves. His model for computing non-parametric conditional quantile functions takes the following form

$$y = f(x) + s(x)e.$$

[He(1997)] takes a three-step procedure where he determines first the conditional median function and then in a second step estimate the smooth non-negative amplitude function. The third step consists of the step wise calculation of the "asymmetry factor" c_α for each $\alpha-$quantile curve separately.

2.4 Expectile and quantile estimation using boosting

1. Initialize all model components as $\hat{\boldsymbol{f}}_j^{[0]}(\boldsymbol{z}) \equiv \mathbf{0}$, $j = 1, \ldots, r$. Set the iteration index to $m = 1$.

2. Compute the current negative gradient vector $\boldsymbol{u}$ with elements

$$u_i = -\frac{\partial}{\partial \eta} \rho(y_i, \eta) \bigg|_{\eta = \hat{\eta}^{[m-1]}(\boldsymbol{z}_i)}, \quad i = 1, \ldots, n.$$

3. Choose the base-learner $\boldsymbol{g}_{j^*}$ that minimizes the L_2-loss, i.e. the best-fitting function according to

$$j^* = \underset{1 \le j \le r}{\arg\min} \sum_{i=1}^{n} (u_i - \hat{g}_j(\boldsymbol{z})_i)^2$$

   where $\hat{\boldsymbol{g}}_j = \boldsymbol{S}_j \boldsymbol{u}$.

4. Update the corresponding function estimate to $\hat{\boldsymbol{f}}_{j^*}^{[m]} = \hat{\boldsymbol{f}}_{j^*}^{[m-1]} + \nu \hat{\boldsymbol{g}}_{j^*}$, where $\nu \in (0, 1]$ is a step size. For all remaining functions set $\hat{\boldsymbol{f}}_j^{[m]} = \hat{\boldsymbol{f}}_j^{[m-1]}$, $j \neq j^*$.

5. Increase m by one. If $m < m_{\text{stop}}$ go back to step 2., otherwise terminate the algorithm.

For expectile regression, the empirical risk is given the asymmetric least squares criterion (1) and the appropriate loss function is defined as $\rho(y, \eta) = w(\tau)(y - \eta_\tau)^2$. The corresponding negative gradient is therefore obtained as

$$u_i = 2w_i(\tau)(y_i - \eta_i).$$

3 Example and available data

Expectile estimation can be used in a almost any type of situation where one is interested in estimating smooth curves in non-central parts of the data under consideration. The data provided with the package are

```
> data(india)
> data(dutchboys)
```

`india` consists of a data sample of 4000 observations with 6 variables from a 'Demographic and Health Survey' about malnutrition of children in India. Data set only contains 1/10 of the observations and some basic variables to enable first analyses. Details are given in [Fenske *et al.*(2009)].
`dutchboys` contains data from the Fourth Dutch growth study and includes 6848 observations on 10 variables. More information can be found in [Van Buuren and Fredriks(2001)].

3.1 Basic examples

The basic function `expectreg.ls` can be used to estimate 11 expectiles curves for different levels of asymmetry parameter p. The results are shown in the following graph.

```
> data(dutchboys)
```

```
> exp.l <- expectreg.ls(dutchboys[,3] ~ rb(dutchboys[,2],"pspline"),smooth="acv")
```

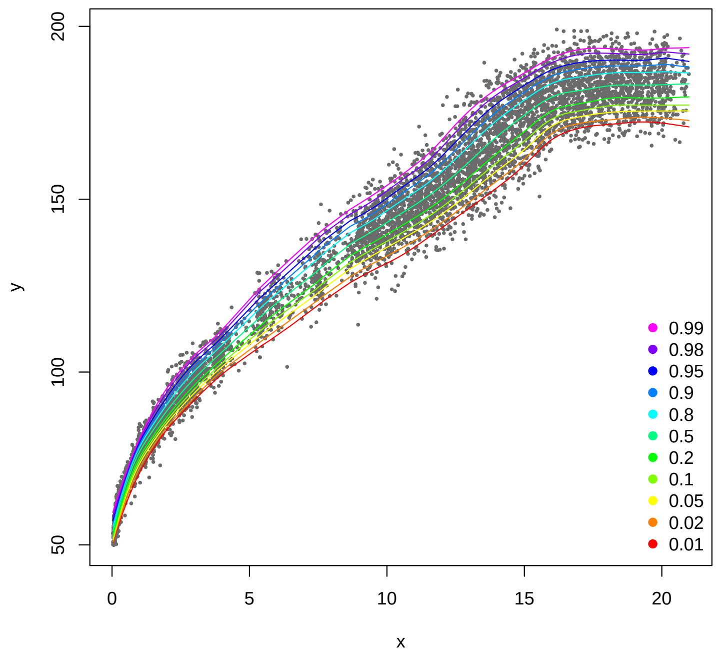

Figure 1: Expectile curves estimated using `expectreg.ls`

Due to the large number of observations in the data set crossing of curves is already unlikely to happen. Nevertheless we apply also the expectile bundle model implemented in `expectile.bundle` to this example.

```
> exp.b <- expectreg.ls(dutchboys[,3] ~ rb(dutchboys[,2],"pspline"),smooth="none",estimate
```

Additionally we analyze the data with the algorithm proposed in [He(1997)] implemented in `expectile.restricted`.

```
> exp.r <- expectreg.ls(dutchboys[,3] ~ rb(dutchboys[,2],"pspline"),smooth="schall",estima
```

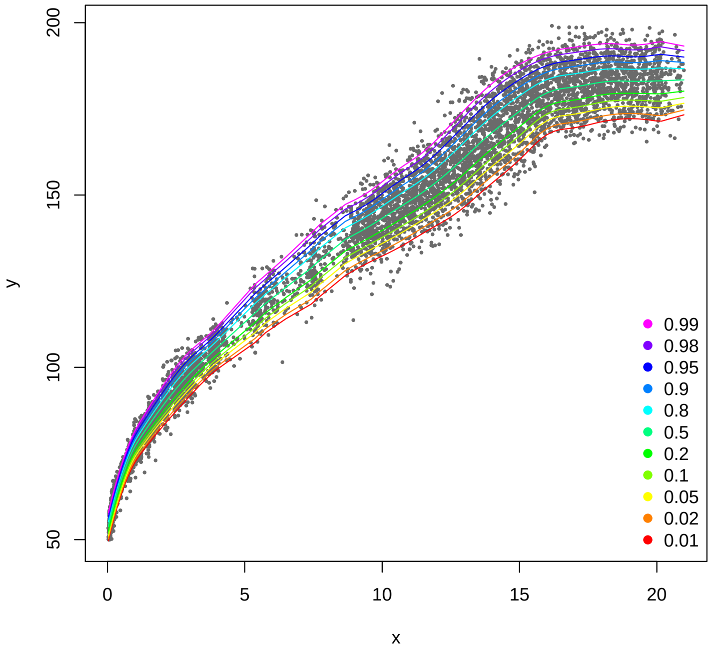

Figure 2: Expectile curves estimated using `expectreg.ls` with bundle estimate

3.2 Applied boosting

```
> exp.boost <- expectreg.boost(hgt ~ bbs(age,df=5,degree=2),dutchboys,mstop=rep(500,11))
```

References

[Van Buuren and Fredriks(2001)] Van Buuren, S., and A. M. Fredriks, 2001 Worm plot: A simple diagnostic device for modeling growth reference curves. Statistics in Medicine **20**: 1259–1277.

[Fenske *et al.*(2009)] Fenske, N., T. Kneib, and T. Hothorn, 2009 Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Techical Report 52, University of Munich.

[He(1997)] He, X., 1997 Quantile curves without crossing. The American Statistician **51**: 186–192.

[Koenker(1992)] Koenker, R., 1992 When are expectiles percentiles? (solution). Economic Theory **9**: 526–527.

[Newey and Powell(1987)] Newey, W. K., and J. L. Powell, 1987 Asymmetric least squares estimation and testing. Econometrica **55**: 819–847.

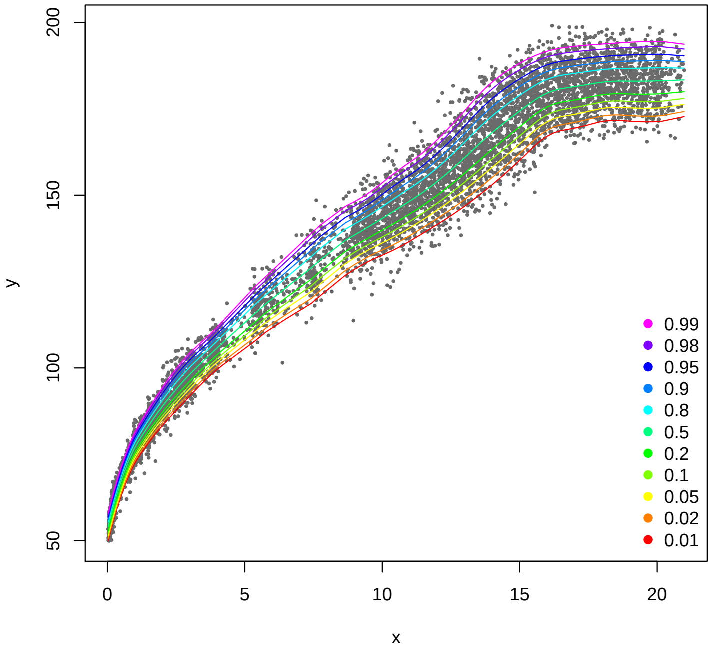

Figure 3: Expectile curves estimated using `expectreg.ls` with restricted estimate

[Schnabel and Eilers(2009)] Schnabel, S. K., and P. H. C. Eilers, 2009b Optimal expectile smoothing. Computational Statistics and Data Analysis **53**: 4168–4177.

[Schnabel and Eilers(2010)] Schnabel, S. K., and P. H. C. Eilers, 2010 Non crossing expectiles and quantiles. Journal for Computational and Graphical Statistics (Submitted).

[Sobotka and Kneib(2010)] Sobotka, F., and T. Kneib, 2010 Geoadditive Expectile Regression. Computational Statistics and Data Analysis, doi: 10.1016/j.csda.2010.11.015.

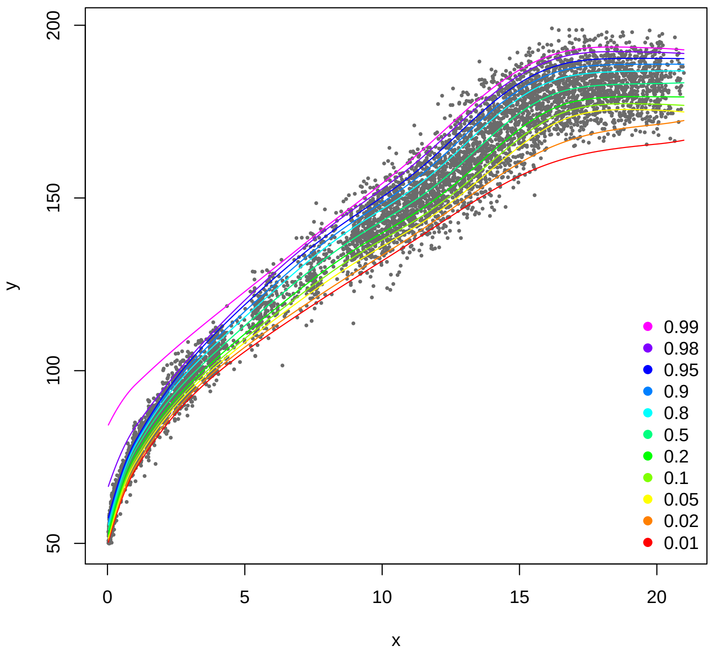

Figure 4: Expectile curves estimated using `expectreg.boost`