GET: Quantile regression

Mari Myllymäki and Mikko Kuronen
Natural Resources Institute Finland (Luke)

Abstract

This vignette gives examples of global quantile regression, as proposed in Mrkvička, Konstantinou, Kuronen, and Myllymäki (2023) and as implemented in the R package **GET**. When citing the vignette and package please cite **?** and Mrkvička *et al.* (2023), and further relevant references given by typing `citation("GET")` in R.

Keywords: global envelope test, goodness-of-fit, Monte Carlo test, R, quantile regression.

1. Introduction

This vignette gives examples of the use of global quantile regression (Mrkvička *et al.* 2023). The examples utilize the R (R Core Team 2023) package **quantreg** (Koenker 2023) in addition to the **GET** package (**?**). The plots are produced by the use of the **ggplot2** package (Wickham 2016), where we utilize the theme `theme_bw` for this document.

```
R> library("GET")
R> library("quantreg")
R> library("ggplot2")
R> theme_set(theme_bw(base_size = 9))
```

2. Data

The **GET** package contains *simulated* data which mimics the example of distribution comparison of natural, near-natural and non-natural forests of Mrkvička *et al.* (2023) (see also Myllymäki, Tuominen, Kuronen, Packalen, and Kangas 2023). The simulated data is available in the data object `naturalness`. The data contains simulated stand ages in the three groups of categorical variables `Naturalness` and `DominantSpecies`.

```
R> data("naturalness")
R> str(naturalness)
```

```
'data.frame':        773 obs. of  3 variables:
 $ DominantSpecies: Factor w/ 3 levels "Conifer","Mixed",..: 3 1 2 1 1 1 3 2 3 1 ...
 $ Naturalness    : Factor w/ 3 levels "Non-natural",..: 1 3 1 1 1 1 1 1 1 1 ...
 $ Age            : int  37 223 64 68 82 68 48 58 33 67 ...
```

We use this simulated data to show the workflow of global quantile regression. Let us specify the quantiles, which we will inspect below. We use 100 quantiles, which are equidistant from each other, with the smallest quantile equal to 0.051 and the largest quantile equal to 0.949.

The number of permutations we set to 2499. For experimenting only, you may like to use a smaller value though.

```
R> taus <- seq(0.051, 0.949, length=100)
R> nperm <- 2499
```

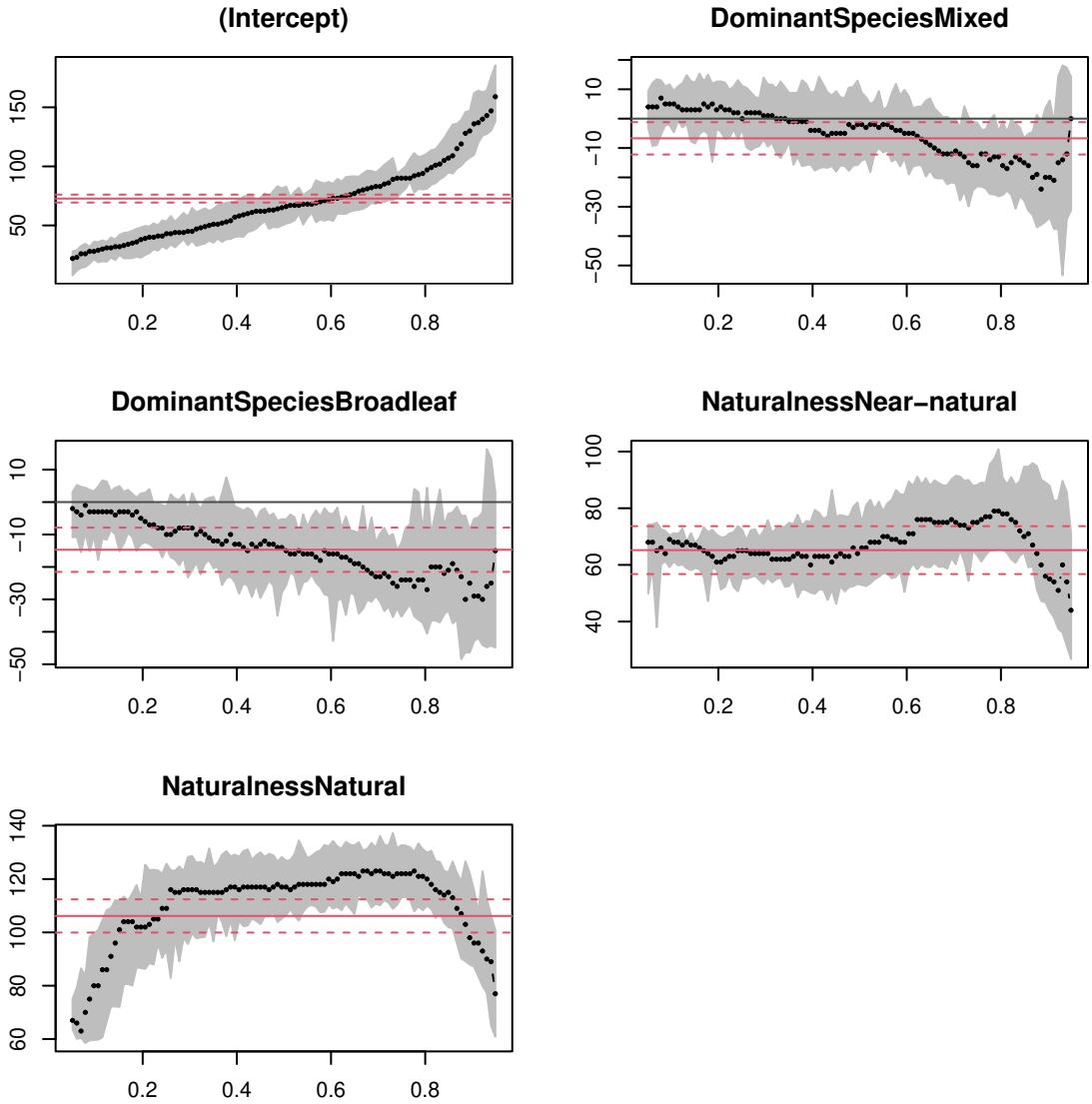
3. Quantile regression

Our interest is first in the naturalness, while the dominant species is the nuisance. The quantile regression model is

$$Age \sim constant + naturalness + species.$$

We first fit the quantile regression model. We use the `rq()` to fit the quantile regression model for each quantile, and the function `summary()` to compute the 95% pointwise confidence intervals.

```
R> r1 <- rq(Age ~ DominantSpecies + Naturalness, data=naturalness, tau = taus)
R> s1 <- summary(r1)
R> plot(s1)
```

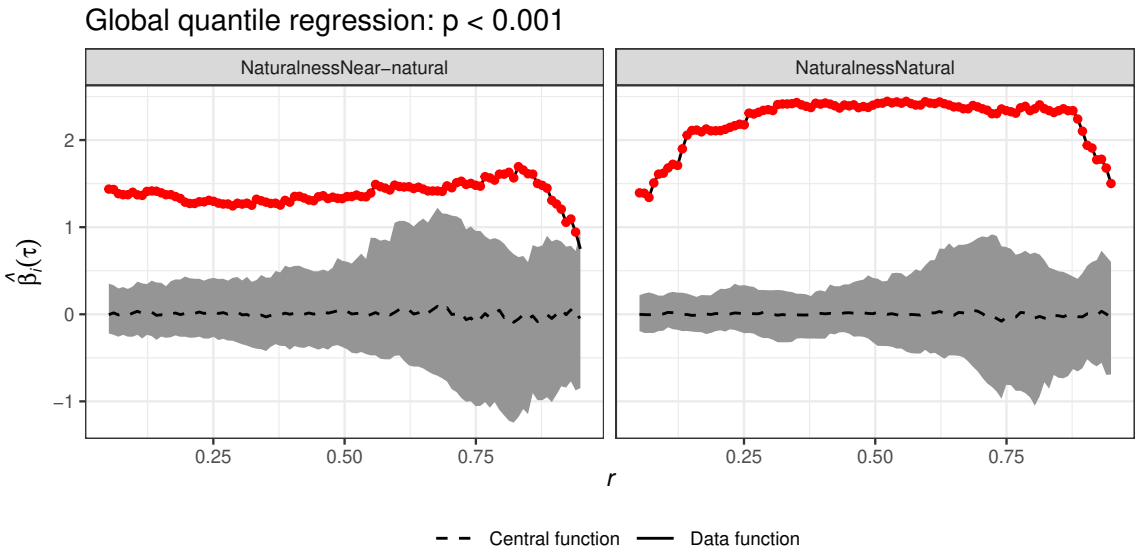


According to the quantile regression fit, the effect of the nuisance effect, i.e., dominant species, appears to be location-scale shift, since the estimated coefficients of the mixed and broadleaf stands appear to be linear in τ. Therefore, to test for the differences between the distributions of stand age in the natural, near-natural and non-natural forests, we choose the permutation algorithm "remove location scale" (RLS) (Mrkvička *et al.* 2023). We use the function `global_rq()` for the global quantile regression to test the difference between the three naturalness groups.

```
R> res <- global_rq(nperm,
+           formula.full = Age ~ DominantSpecies + Naturalness,
+           formula.reduced = Age ~ DominantSpecies,
+           data = naturalness,
+           typeone = "fwer",
+           permutationstrategy = "remove location scale",
+           taus = taus)
```

We can directly plot the result.

```
R> plot(res)
```



The grey zone shows the global envelope, while the estimated coefficients are shown by a black solid line, overlaid with red dots when outside the envelope. Note here that the global test of naturalness contains both functional coefficients shown in the plot. The test identifies both the significant quantiles and the corresponding coefficient under the global test. Here, the coefficients of near-natural and natural stands show the difference to non-natural reference group for all quantiles: both the near-natural and natural stands are uniformly (for all quantiles) older than non-natural stands.

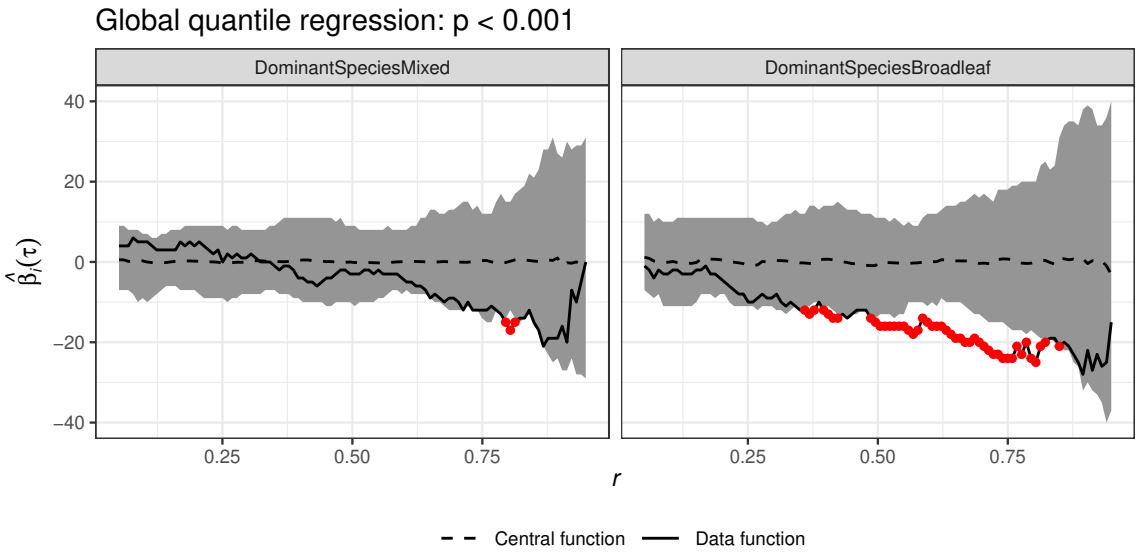
Secondly, we change the role of naturalness and dominant species, keeping now the dominant species as the interesting factor. Because the effect of the naturalness groups does not seem to be linear (see the plot of the quantile regression above), the location-scale shift can not be assumed and we choose the "remove quantile" (RQ) permutation strategy. The test for the effect of the dominant species is as follows:

```
R> res2 <- global_rq(nperm,
+           formula.full = Age ~ DominantSpecies + Naturalness,
+           formula.reduced = Age ~ Naturalness,
+           data = naturalness,
+           typeone = "fwer",
+           permutationstrategy = "remove quantile",
+           taus = taus)
R> res2

Global quantile regression (one-step) (1d):
 * Based on the measure: "erl"
 * 95% global envelope
 * p-value of the global test: 4e-04
```

```
 * Significance level of the global test: 0.05
The object contains a list of 2 components
 * each containing: $r $obs $central $lo $hi
 * Number of r-values with observed function outside the envelope: 3 46
 * Total number of argument values r                              : 100 100
```

```
R> plot(res2)
```



Here we observe differences between the groups for specific quantiles. The negative coefficient suggest that the stand age distribution of broadleaf dominated forests is more skewed to the left than the distribution of conifer dominated forests, but the ranges of stand ages are equal in the different categories. The difference of mixed and broadleaf stands is similar, but the significant effect occurs for smaller number of quantiles.

References

Koenker R (2023). ***quantreg**: Quantile Regression*. R package version 5.97, URL https://CRAN.R-project.org/package=quantreg.

Mrkvička T, Konstantinou K, Kuronen M, Myllymäki M (2023). "Global quantile regression." arXiv:2309.04746 [stat.ME]. doi:10.48550/arXiv.2309.04746.

Myllymäki M, Tuominen S, Kuronen M, Packalen P, Kangas A (2023). "The relationship between forest structure and naturalness in the Finnish national forest inventory." *Forestry: An International Journal of Forest Research*, p. cpad053. doi:10.1093/forestry/cpad053.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. ISBN 978-3-319-24277-4.

Affiliation:

Mari Myllymäki
Natural Resources Institute Finland (Luke)
Latokartanonkaari 9
FI-00790 Helsinki, Finland
E-mail: mari.myllymaki@luke.fi
URL: https://www.luke.fi/en/experts/mari-myllymaki/
and
Mikko Kuronen
Natural Resources Institute Finland (Luke)
Latokartanonkaari 9
FI-00790 Helsinki, Finland
E-mail: mikko.kuronen@luke.fi
URL: https://www.luke.fi/en/experts/mikko-kuronen/