Automated Model Building and Goodness-of-fit via Quantile Regression

Haim Bar

September 5, 2022

Abstract

This repository contains code and data used in the paper *Automated Model Building and Goodness-of-fit via Quantile Regression* by Bar, Booth, and Wells. Given P predictors x_i and n observations for each x_i and the response variable y, the goal is to build a model, $y = f(x_1, \dots, x_P)$ where $f()$ consists of combinations of powers of the x_i's, which fits the data well across multiple quantiles.

1 Prerequisites

In order to run the code you must first install the **QREM** package. Since **QREM** has a model selection option for cases in which the number of predictors is large you also need to install the packages **edgefinder** and **SEMMS**:

```
devtools::install_github("haimbar/edgefinder")
devtools::install_github("haimbar/SEMMS")
devtools::install_github("haimbar/QREM")
```

The model building algorithm is implemented in a function called *fitQRloop* in the file runQREM.R. The function takes five arguments:

- M the data matrix with P columns and n rows.
- qns The quantile which will be used in the fitting algorithm.
- minDiff The minimal improvement in the overall goodness of fit in order to accept a new term.
- maxdeg The maximum degree of any term in the model.
- maxrows The maximum number of rows in the matrix of possible terms up to degree maxdeg.

The file initSim.R contains the values we used by default. It also contains three other variables which are used by **QREM** in the fitting process:

- mxm The maximum number of segments in the partition of the selected variable.
- alphaQ The level of the goodness of fit test.
- plotit A Boolean variable which tells the function *flatQQplot* whether to show intermediate diagnostic plots for each accepted new term in the model.

```
qns <- 1:5/6
k <- length(qns)
minDiff <- 4
maxdeg <- 15
maxrows <- 5000
mxm <- 30
```

```
alphaQ <- 0.01
plotit <- FALSE
```

2 A Univariate Example

The file Code/Univariate02.R contains the code for example #1 in the paper:

```
N <- 5000
set.seed(211111)
x <- runif(N, min=0, max=4*pi)
y <- exp(-x)*x^5 + rnorm(N, 0, 0.25*(x+0.05)) # EXAMPLE 1 in the paper
M <- data.frame(y=y, x1=x)
res <- fitQRloop(M=M, qn = qns, maxdeg = maxdeg, minDiff = minDiff)
pdf("Figures/Uni02.pdf", width=5, height=5)
plot(x, y, cex=0.5, pch=19, col="grey66", axes=F)
axis(1); axis(2); grid()
for (i in 1:k) {
  lines(sort(x), res$qremFit[[i]]$fitted.mod$fitted.values[order(x)],
   ↪  col=2)
}
```

?? is a comment

References

[1] Bar, H. Y., Booth, J. G., and Wells, M. T. (2020). A Scalable Empirical Bayes

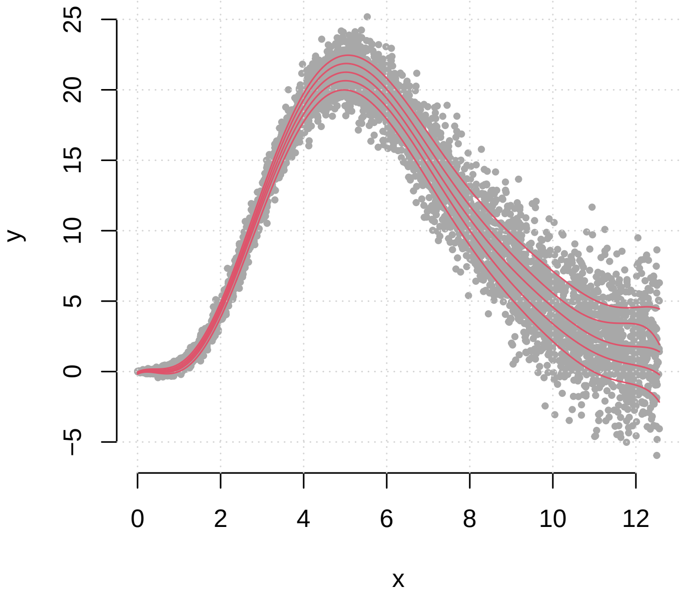

Figure 1: Simulation 23 – Diagnostic plot using the `QRdiagnostics` function. The true model is $y \sim N(6x^2 + x + 120, (0.1 + 0.5x)^2))$

Approach to Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics*, **0**(0), 1–12.