Quantile regression robustness

Anton Antonov
Mathematica for Prediction at GitHub
Mathematica for Prediction blog
December 2013

Introduction

This document shows examples of robustness of quantile regression. The quantile regression curves (called "regression quantiles") are computed with the *Mathematica* package [1] hosted at the MathematicaForPrediction project at GitHub. Quantile regression was introduced by Koenker and Bassett in [3] and detailed theoretical descriptions and discussions are given the book "Quantile regression" by Koenker (see [4]).

This document extends the descriptions in [2]. More complicated data distributions can be used, but the data used in the document although based on a simple deterministic model has a simple heteroscedasticity (the variance varies with *x*, see [5]).

Load

Load the package [1]:

In[1575]:=
```
Get["~/MathFiles/MathematicaForPrediction/QuantileRegression.m"]
```

Here is the definition of the data generation function used below:

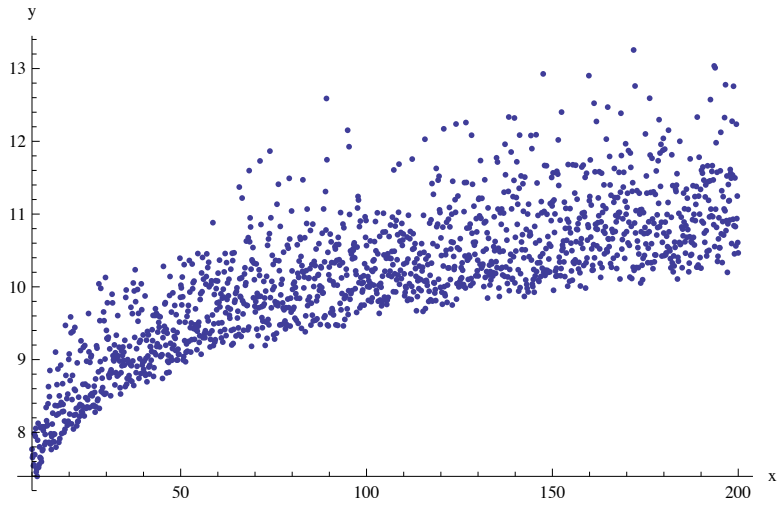
In[1577]:=
```
Clear[LogarithmicCurveWithNoise]
LogarithmicCurveWithNoise[
   nPoints_Integer, start_?NumberQ, end_?NumberQ] :=
  Block[{data},
   data =
    Table[{t, 5 + Log[t] + RandomReal[SkewNormalDistribution[0, Log[t]/5, 12]]},
     {t, Rescale[Range[1, nPoints], {1, nPoints}, {start, end}]}];
   data
  ];
```

Starting data and regression quantiles

Let us generate some data.

In[1579]:=
```
data = LogarithmicCurveWithNoise[1500, 10, 200];
ListPlot[data, AxesLabel → {"x", "y"}, PlotRange → All, ImageSize → 400]
```

Out[1580]=



We know that the model for the data is

$$y = \beta_0 + \beta_1 x + \beta_3 \log(x), \tag{1}$$

hence are going to fit the model functions:

In[1581]:=
```
funcs = {1, x, Log[x]};
```

Here are the regression quantiles:

In[1582]:=
```
qrFuncs = QuantileRegression[data, funcs, x, qs];
TableForm[List /@ Chop[qrFuncs]]
```

Out[1583]//TableForm=

$5.04313 + 1.00043 \text{ Log}[x]$

$5.01161 + 1.06001 \text{ Log}[x]$

$5.04302 + 8.62724 \times 10^{-10}\, x + 1.12171 \text{ Log}[x]$

$5.08776 + 1.20479 \text{ Log}[x]$

$5.73968 + 0.00223455\, x + 1.16063 \text{ Log}[x]$

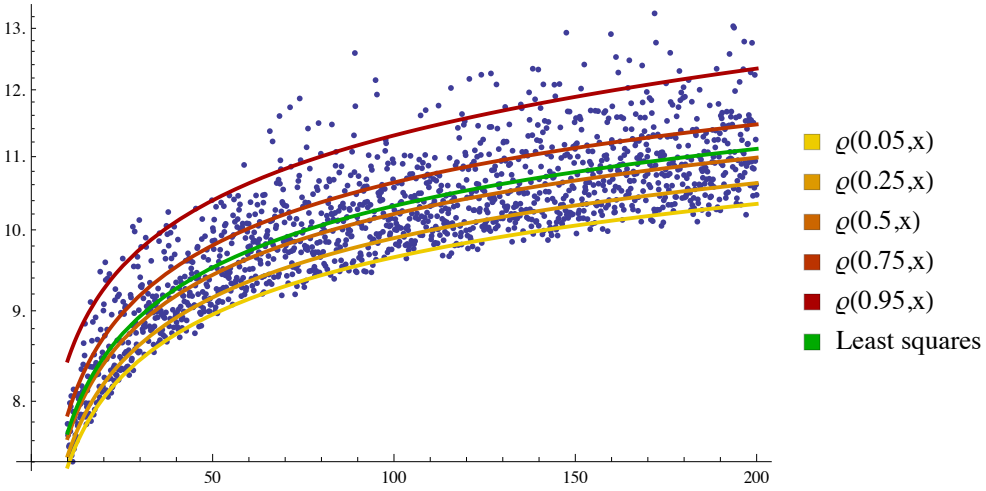
Here is the least regression fit:

In[1584]:=
```
fFunc = Fit[data, funcs, x]
```

Out[1584]= $5.04503 + 0.0000291599\, x + 1.14402 \text{ Log}[x]$

Here is a plot of the data and all fitted functions:

Out[1586]=

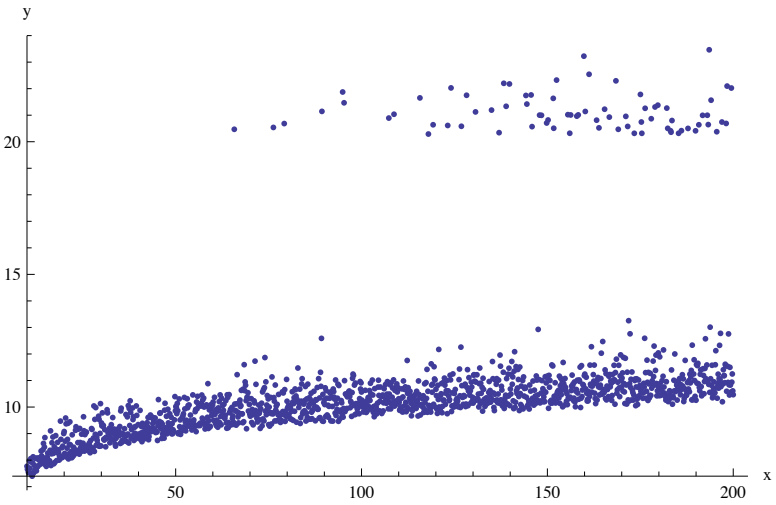


Outliers above a constant

Let us demonstrate the robustness of the regression quantiles with the data of the previous example. Suppose that for some reason 50% of the data *y*-values greater than 11.25 are altered by multiplying them with a some greater than 1 factor, say, $\alpha = 1.8$. Then the altered data looks like this:

In[1588]:=
```
α = 1.8;
dataAlt =
  Map[If[RandomInteger[{0, 1}] == 1 && #[[2]] > 11.25, {#[[1]], α #[[2]]}, #] &, data];
ListPlot[dataAlt, AxesLabel → {"x", "y"}, PlotRange → All, ImageSize → 400]
```

Out[1590]=



Let us compute the regression quantiles for the altered data:

In[1591]:=
```
qrFuncsAlt = QuantileRegression[dataAlt, funcs, x, qs];
TableForm[List /@ Chop[qrFuncs]]
```

Out[1592]//TableForm=

$5.04313 + 1.00043 \text{ Log}[x]$

$5.01161 + 1.06001 \text{ Log}[x]$

$5.04302 + 8.62724 \times 10^{-10} x + 1.12171 \text{ Log}[x]$

$5.08776 + 1.20479 \text{ Log}[x]$

$5.73968 + 0.00223455 x + 1.16063 \text{ Log}[x]$

and let us also compute the least squares fit of the model (1):

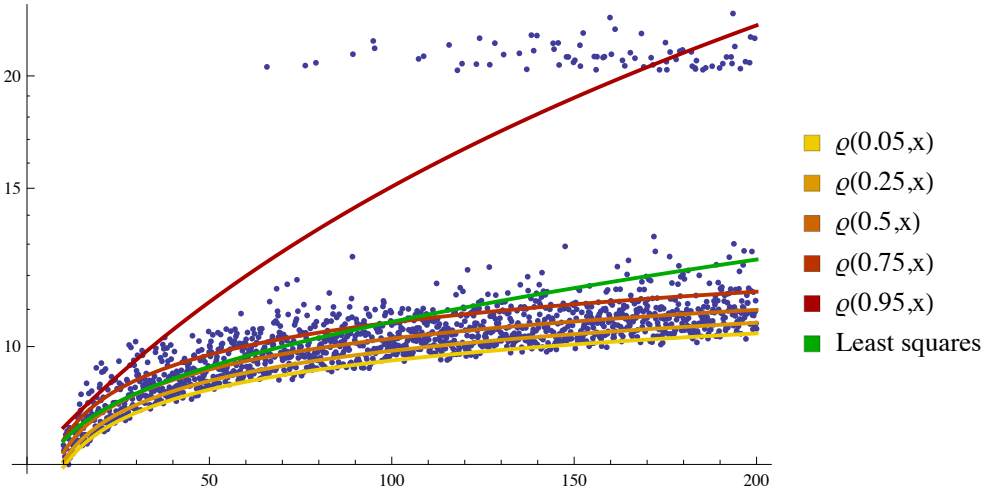
In[1593]:=
```
fFunc = Fit[data, funcs, x]
fFuncAlt = Fit[dataAlt, funcs, x]
```

Out[1593]= $5.04503 + 0.0000291599 x + 1.14402 \text{ Log}[x]$

Out[1594]= $6.14823 + 0.0137224 x + 0.680375 \text{ Log}[x]$

Here is a plot of the altered data and all fitted functions:

Out[1596]=



We can see that the new regression quantiles computed for 0.05, 0.25, and 0.5 have not changed significantly:

In[1598]:=
```
Grid[List /@ Flatten[Transpose[{Chop[qrFuncs], Chop[qrFuncsAlt]}]],
 Alignment → Left, Dividers →
  {{True, True}, Flatten@{Table[{True, False}, {Length[qrFuncs]}], True}}]
```

Out[1598]=

$5.04313 + 1.00043 \text{ Log}[x]$ $5.04313 + 1.00043 \text{ Log}[x]$
$5.01161 + 1.06001 \text{ Log}[x]$ $5.01161 + 1.06001 \text{ Log}[x]$
$5.04302 + 8.62724 \times 10^{-10} x + 1.12171 \text{ Log}[x]$ $5.04302 + 2.33272 \times 10^{-10} x + 1.12171 \text{ Log}[x]$
$5.08776 + 1.20479 \text{ Log}[x]$ $5.1547 + 0.000541577 x + 1.1789 \text{ Log}[x]$
$5.73968 + 0.00223455 x + 1.16063 \text{ Log}[x]$ $7.35883 + 0.0770217 x$

ant that they are still good for separating the un-altered data:

In[1599]:=
```
tbl = Table[{qs[[i]], Length[Select[data, #[[2]] ≥ (qrFuncsAlt[[i]] /. x → #[[1]]) &]] /
      Length[data] // N}, {i, Length[qs]}];
TableForm[tbl, TableHeadings → {None, {"quantile", "fraction\nabove"}}]
```

Out[1599]//TableForm=

quantile	fraction above
0.05	0.95
0.25	0.750667
0.5	0.5
0.75	0.242667
0.95	0.0153333

Also we can see that the least squares fit of (1) has significantly changed:

In[1600]:=
```
fFunc
```

Out[1600]=
```
5.04503 + 0.0000291599 x + 1.14402 Log[x]
```

In[1601]:=
```
fFuncAlt
```

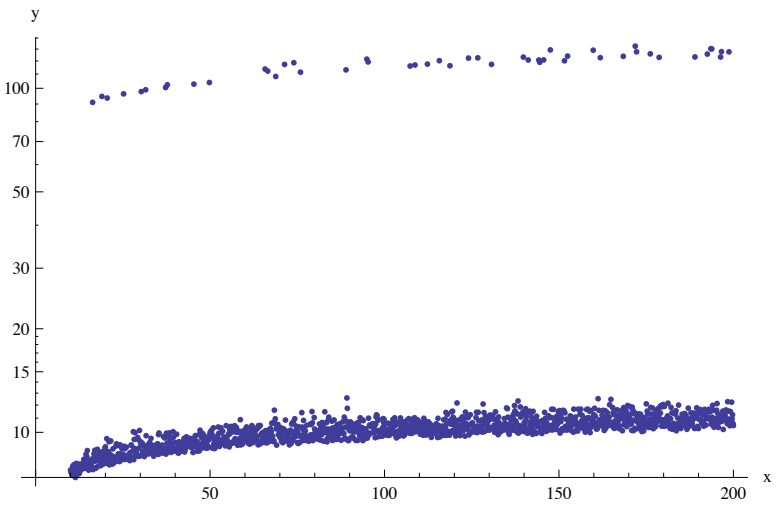
Out[1601]=
```
6.14823 + 0.0137224 x + 0.680375 Log[x]
```

Outliers above a chosen regression quantile

Let us demonstrate the robustness of the regression quantiles with different type of data alternation. Suppose that for some reason 70 % of the data *y*-values above the altered by multiplying them with a some greater than 1 factor, say, $\alpha = 10$. Then the altered data looks like this:

In[1602]:=
```
α = 10.;
dataAlt = Map[If[RandomReal[{0, 1}] > 0.3 && (#[[2]] > (qrFuncs[[5]] /. x -> #[[1]])),
      {#[[1]], α #[[2]]}, #] &, data];
ListLogPlot[dataAlt, AxesLabel → {"x", "y"}, PlotRange → All, ImageSize → 400]
```

Out[1604]=



Let us compute the regression quantiles for the altered data:

In[1605]:=
```
qrFuncsAlt = QuantileRegression[dataAlt, funcs, x, qs];
TableForm[List /@ Chop[qrFuncs]]
```

Out[1606]//TableForm=

$5.04313 + 1.00043 \text{ Log}[x]$

$5.01161 + 1.06001 \text{ Log}[x]$

$5.04302 + 8.62724 \times 10^{-10} x + 1.12171 \text{ Log}[x]$

$5.08776 + 1.20479 \text{ Log}[x]$

$5.73968 + 0.00223455 x + 1.16063 \text{ Log}[x]$

and let us also compute the least squares fit of the model (1):

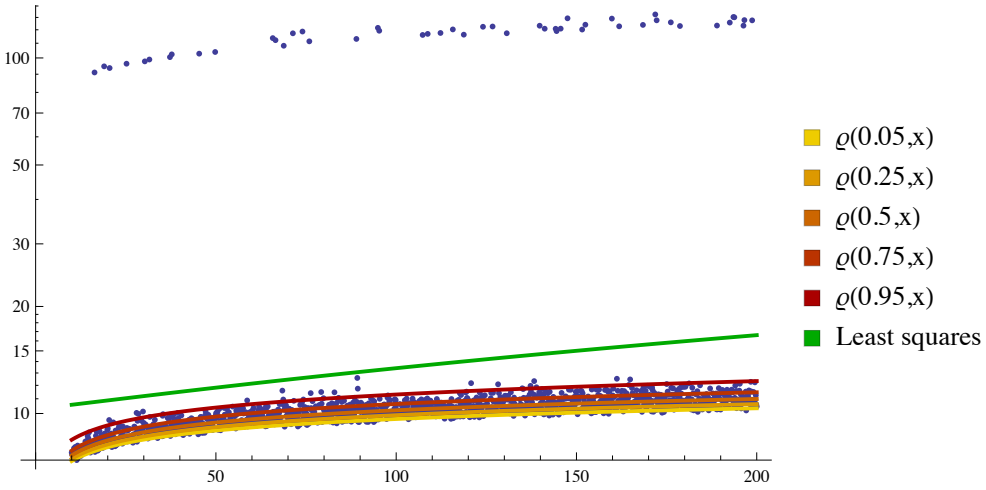
In[1607]:=
```
fFunc = Fit[data, funcs, x]
fFuncAlt = Fit[dataAlt, funcs, x]
```

Out[1607]= $5.04503 + 0.0000291599 x + 1.14402 \text{ Log}[x]$

Out[1608]= $10.3294 + 0.0322001 x - 0.0301403 \text{ Log}[x]$

Here is a plot of the altered data and all fitted functions:

Out[1610]=



We can see that the new regression quantiles computed for 0.05, 0.25, 0.5, 0.75, and 0.95 have not changed significantly:

In[1612]:=
```
Grid[List /@ Flatten[Transpose[{Chop[qrFuncs], Chop[qrFuncsAlt]}]],
 Alignment → Left, Dividers →
  {{True, True}, Flatten@{Table[{True, False}, {Length[qrFuncs]}], True}}]
```

Out[1612]=

$5.04313 + 1.00043 \text{ Log}[x]$	$5.04313 + 1.00043 \text{ Log}[x]$
$5.01161 + 1.06001 \text{ Log}[x]$	$5.01161 + 1.68771 \times 10^{-8} x + 1.06001 \text{ Log}[x]$
$5.04302 + 8.62724 \times 10^{-10} x + 1.12171 \text{ Log}[x]$	$5.04302 + 2.78614 \times 10^{-9} x + 1.12171 \text{ Log}[x]$
$5.08776 + 1.20479 \text{ Log}[x]$	$5.08776 + 1.81456 \times 10^{-10} x + 1.20479 \text{ Log}[x]$
$5.73968 + 0.00223455 x + 1.16063 \text{ Log}[x]$	$5.73912 + 0.00223246 x + 1.16081 \text{ Log}[x]$

ant that they are still good for separating the un-altered data:

```
In[1613]:= tbl = Table[{qs[[i]], Length[Select[data, #[[2]] ≥ (qrFuncsAlt[[i]] /. x → #[[1]]) &]] /
        Length[data] // N}, {i, Length[qs]}];
      TableForm[tbl, TableHeadings → {None, {"quantile", "fraction\nabove"}}]
```

Out[1613]//TableForm=

quantile	fraction above
0.05	0.950667
0.25	0.750667
0.5	0.499333
0.75	0.25
0.95	0.0493333

Also we can see that the least squares fit of (1) has significantly changed:

```
In[1614]:= fFunc
Out[1614]= 5.04503 + 0.0000291599 x + 1.14402 Log[x]

In[1615]:= fFuncAlt
Out[1615]= 10.3294 + 0.0322001 x - 0.0301403 Log[x]
```

Conclusions

The examples considerd clearly demonstrate the robustness of quantile regression when compared to the least squares fit. As in the single distribution case, computing quantiles can be very useful for identifying outliers. For example, we can do the regression analogue of standardizing the data by subtracting the median and dividing by the interquartile distances, and declare any point outside of specified range as an outlier.

References

[1] Anton Antonov, Quantile regression *Mathematica* package, source code at GitHub, https://github.-com/antononcube/MathematicaForPrediction, package QuantileRegression.m, (2013).

[2] Anton Antonov, “Quantile regeression through linear programming”, a package usage guide at GitHub, https://github.com/antononcube/MathematicaForPrediction, (2013).
URL: https://github.com/antononcube/MathematicaForPrediction/blob/master/Documentation/Quantile%20 regression %20 through %20 linear %20 programming.pdf .

[3] Roger Koenker, Gilbert Bassett Jr., “Regression Quantiles”, Econometrica, 46(1), 1978, pp. 33-50. JSTOR URL: http://links.jstor.org/sici?sici=0012-9682%28197801 %2946 %3 A1 %3 C33 %3 ARQ %3 E2 .0.CO%3 B2-J .

[4] Roger Koenker, Quantile regression, Econometric Society Monographs (Book 38), Cambridge University Press, (2005). Google URL: http://books.google.com/books/about/Quantile_Regression.html?id=hdkt7V4NXsgC .

[5] Wikipedia entry, Heteroscedasticity, http://en.wikipedia.org/wiki/Heteroscedasticity .