Quantile Regression
LSE Short Course: 16-17 May 2011[1]

Roger Koenker
CEMMAP and University of Illinois, Urbana-Champaign

Quantile regresson extends classical least squares methods for estimating conditional mean functions by offering a variety of methods for estimating conditional quantile functions, thereby enabling the researcher to explore more thoroughly heterogeneous covariate effects. The course will offer a comprehensive introduction to quantile regression methods and briefly survey some recent developments. The primary reference for the course will be my 2005 Econometric Society monograph, but further readings are suggested below in this course outline.

Course lectures will be complemented by several computationally oriented interludes designed to give students some experience with applications of the methods. These sessions will be conducted in the opensource R language, and will rely heavily on my `quantreg` package. Thus it would be helpful if students brought laptops equipped with this software already installed. R can be freely downloaded for PC/Mac/Linux machines from CRAN: `http://cran.r-project.org/`. The `quantreg` package is also available from CRAN, just click on "packages" on the left margin of the page and follow the directions you will find there. Students familiar with Stata and wanting to experiment with Stata data sets should consider also downloading the "foreign" package, which contains a function called `read.dta` that enables R to read Stata data files.

Tentative Topics

(1) The Basics: What, Why and How? Koenker (2005, §1-2), Koenker and Hallock (2001)
(2) Inference and Quantile Treatment Effects Koenker (2005, §3),
(3) Nonparametric Quantile Regression Koenker (2005, §7), Koenker (2010),Belloni and Chernozhukov (2009)
(4) Endogoneity and IV Methods Chesher (2003) Chernozhukov and Hansen (2005) Ma and Koenker (2005)
(5) Censored QR and Survival Analysis Koenker and Geling (2001) Portnoy (2003) Peng and Huang (2008) Koenker (2008)

[1]Version: May 13, 2011.

(6) Quantile Autoregression Koenker and Xiao (2006)
(7) QR for Longitudinal Data Koenker (2004) Galvao (2009)
(8) Risk Assessment and Choquet Portfolios Bassett, Koenker, and Kordas (2004)
(9) Quantile Regression Computation: From the Inside and Outside Koenker (2005, §6),

References

Bassett, G., R. Koenker, and G. Kordas (2004): "Pessimistic Portfolio Allocation and Choquet Expected Utility," *J. of Financial Econometrics*, 2, 477–92.
Belloni, A., and V. Chernozhukov (2009): "L_1-Penalized Quantile Regression in High-Dimensional Sparse Models," forthcoming Annals of Statistics.
Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.
Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.
Galvao, A. (2009): "Quantile Regression for Dynamic Panel Data with Fixed Effects," forthcoming, *J. of Econometrics*.
Koenker, R. (2004): "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91, 74–89.
——— (2005): *Quantile Regression*. Cambridge.
——— (2008): "Censored Quantile Regression Redux," *J. of Statistical Software*, 27(6), 1–24.
——— (2010): "Additive Models for Quantile Regression: Model Selection and Confidence Bandaids," forthcoming, *Brazilian J. of Statistics*.
Koenker, R., and O. Geling (2001): "Reappraising Medfly Longevity: A quantile regression survival analysis," *J. of Am. Stat. Assoc.*, 96, 458–468.
Koenker, R., and K. Hallock (2001): "Quantile Regression," *J. of Economic Perspectives*, 15, 143–156.
Koenker, R., and Z. Xiao (2006): "Quantile Autoregression, with discussion and rejoinder," *J. of Am. Stat. Assoc.*, 96, 980–1006.
Ma, L., and R. Koenker (2005): "Quantile Regression for Recursive Structural Models," *J. of Econometrics*, 134, 471–506.
Peng, L., and Y. Huang (2008): "Survival Analysis with Quantile Regression Models," *J. of Am. Stat. Assoc.*, 103(482), 637–649.
Portnoy, S. (2003): "Censored Quantile Regression," *J. of Am. Stat. Assoc.*, 98, 1001–1012.

Yet Another R FAQ, or
How I Learned to Stop Worrying and Love Computing [1]

Roger Koenker
CEMMAP and University of Illinois, Urbana-Champaign

> "It was a splendid mind. For if thought is like the keyboard of a piano, divided into so many notes, or like the alphabet is ranged in twenty-six letters all in order, then his splendid mind had no sort of difficulty in running over those letters one by one, firmly and accurately, until it had reached the letter Q. He reached Q. Very few people in the whole of England reach the letter Q.... But after Q? What comes next?... Still, if he could reach R it would be something. Here at least was Q. He dug his heels in at Q. Q he was sure of. Q he could demonstrate. If Q then is Q–R–.... Then R... He braced himself. He clenched himself.... In that flash of darkness he heard people saying–he was a failure–that R was beyond him. He would never reach R. On to R, once more. R—....
> ...He had not genius; he had no claim to that: but he had, or he might have had, the power to repeat every letter of the alphabet from A to Z accurately in order. Meanwhile, he stuck at Q. On then, on to R."
>
> Virginia Woolf (To the Lighthouse)

1. How to get it? Google CRAN, click on your OS, and download. Buy a case of wine with what you've saved.
2. How to start? Click on the R icon if you are mousey, type R in a terminal window if you are penguinesque.
3. What next? At the prompt, > type `2 + 2`
4. What next? At the prompt, > type `1:9/10`
5. What next? At the prompt, > type `x <- 1:99/100`

[1] Version: May 13, 2011. Prepared for an LSE Short Course on Quantile Regression: 16-17 May 2011. More official R FAQs are available from the CRAN website. A FAQ for the quantile regression package **quantreg** can be found by the invoking the command `FAQ()` from within R after loading the package.

6. What next? At the prompt, > type `plot(x,sin(1/x))`
7. What next? At the prompt, > type `lines(x,sin(1/x),col = "red")`
8. How to stop? Click on the Stop sign if you are mousey, type `q()` if you are penguinesque.
9. Isn't there more to R? Yes, try downloading some packages: using the menu in the GUI if you are mousey, or typing install.packages("pname") if you are penguinesque.
10. What's a package? A package is a collection of R software that augments in some way the basic functionality of R, that is it is a way of going "beyond R." For example, the **quantreg** package is a collection of functions to do quantile regression. There were 2992 packages on CRAN as of May 13, 2011.
11. How to use a package? Downloading and installing a package isn't enough, you need to tell R that you would like to use it, for this you can either type: `require(pname)` or `library(pname)`. I prefer the former.
12. How to read data files? For simple files with values separated by white space you can use `read.table`, or `read.csv` for data separated by commas, or some other mark. For more exotic files, there is `scan`. And for data files from other statistical environments, there is the package **foreign** which facilitates the reading of Stata, SAS and other data. There are also very useful packages to read html and other files from the web, but this takes us beyond our introductory objective.
13. What is a `data.frame`? A `data.frame` is a collection of related variables; in the simplest case it is simply a data matrix with each row indexing an observation. However, unlike conventional matrices, the columns of a `data.frame` can be non-numeric, e.g. logical or character or in R parlance, "factors." In many R functions one can specify a `data = "dframe"` argument that specifies where to find the variables mentioned elsewhere in the call.
14. How to get help? If you know what command you want to use, but need further details about how to use it, you can get help by typing `?fname`, if you don't know the function name, then you might try `apropos("concept")`. If this fails then a good strategy is to search `http://finzi.psych.upenn.edu/search.html` with some relevant keywords; here you can specify that you would like to search through the R-help newsgroup, which is a rich source of advice about all things R.

15. Are there manuals? Yes, of course there are manuals, but only to be read as a last resort, but when things get desparate you can always RTFM. The left side of the CRAN website has links to manuals , FAQs and contributed documentation. Some of the latter category is quite good, and is also available in a variety of natural languages. There is also an extensive shelf of published material about R, but indulging in this tends to put a crimp in one's wine budget.

16. What about illustrative examples? A strength of R is the fact that most of the documentation files for R functions have example code that can be easily executed. Thus, for example if you would like to see an example of how to use the command `rq` in the **quantreg** package, you can type `example(rq)` and you will see some examples of its use. Alternatively, you can cut and paste bits of the documentation into the R window; in the OSX GUI you can simply highlight code in a help document, or other window and then press Command-Enter to execute.

17. What's in a name? Objects in R can be of various types called classes. You can create objects by assignment, typically as above with a command like `f <- function(x,y,z)`. A list of the objects currently in your private environment can be viewed with `ls()`, objects in lower level environments like those of the packages that you have loaded can be viewed with `ls(k)` where `k` designates the number of the environment. A list of these environments can be seen with `search()`. Objects can be viewed by simply typing their name, but sometimes objects can be very complicated so a useful abbreviated summary can be obtained with `str(object)`.

18. What about my beloved least squares? Fitting linear models in R is like taking a breath of fresh air after inhaling the smog of other industrial environments. To do so, you specify a model `formula` like this: `lm(y ~ x1 + x2 + x3, data = "dframe")`, if one or more of the x's are factor variables, that is take discrete, qualitative values, then they are automatically exanded into several indicator variables. Interactions plus main effects can be specified by replacing the "+" in the formula by "*". Generalized linear models can be specified in much the same way, as can quantile regression models using the **quantreg** package.

19. What about class conflict? Class analysis can get complicated, but you can generally expect that classes behave themselves in accordance with their material conditions. Thus, for example, suppose you have fitted a linear regression model by least squares using the command `f <- lm(y ~ x1 + x2 + x3)`, thereby assigning the fitted object to the symbol `f`. The object `f`

will have class `lm`, and when you invoke the command `summary(f)`, R will try to find a summary method appropriate to objects of class `lm`. In the simplest case this will entail finding the command `summary.lm` which will produce a conventional table of coefficients, standard errors, t-statistics, p-values and other descriptive statistics. Invoking `summary` on a different type of object, say a data.frame, will produce a different type of summary object. Methods for prediction, testing, plotting and other functionalities are also provided on a class specific basis.

20. What about graphics? R has a very extensive graphics capability. Interactive graphics of the type illustrated already above is quite simple and easy to use. For publication quality graphics, there are device drivers for various graphical formats, generally I find that `pdf` is satisfactory. Dynamic and 3D graphics can be accessed from the package **rgl**.

21. Latex tables? The package **Hmisc** has very convenient functions to convert R matrices into latex tables.

22. Random numbers? There is an extensive capability for generating pseudo random numbers from R. Reproducibility of random sequences is ensured by using the `set.seed` command. Various distributions are accessible with families of functions using the prefixes pdqr, thus for example `pnorm`, `dnorm`, `qnorm` and `rnorm` can be used to evaluate the distribution function, density function, quantile function, or to generate random normals, respectively. See `?Distributions` for a complete list of standard distributions available in base R in this form. Special packages provide additional scope, although it is sometimes tricky to find them.

23. Programming and simulation? The usual language constructs for looping, switching and data management are available, as are recent developments for exploiting multicore parallel processing. Particularly convenient are the family of `apply` functions that facilitate summarizing matrix and list objects. A good way to learn the R language is to look at the code for existing functions. Most of this code is easily accessible from the R command line. If you simply type the name of an R function, you will usually be able to see its code on the screen. Sometimes of course, this code will involve calls to lower level languages, and this code would have to be examined in the source files of the system. But everything is eventually accessible. If you don't like the way a function works you can define a modified version of it for your private use. If you are inspired to write lower level code this is also easily incorporated into the language as explained in the manual called "Writing R Extensions."

SOME EXERCISES ON QUANTILE REGRESSION

ROGER KOENKER

Introduction

These exercises are intended to provide an introduction to quantile regression computing and illustrate some econometric applications of quantile regression methods. For purposes of the course my intention would be to encourage all students to do the first exercise, which gives an overview of the quantile regression software in R in the context of an elementary bivariate Engel curve example. The remaining exercises are more open ended. I would like students to choose one of these exercises according to their own special interests. Given the brief duration of the course, it is obviously unrealistic to expect answers to these questions at the end of the course, but I would be happy to get responses via email should you choose to continue working on them after the course is finished.

A Word on Software. There is now some quantile regression functionality in most statistical software systems. Not surprisingly, I have a strong preference for the implementation provide by the **quantreg** package of R, since I've devoted a considerable amount of effort to writing it. R is a dialect of John Chambers's S language and provides a very general, very elegant environment for data analysis and statistical research. It is fair to say that R is now the vehicle of choice within the statistical computing community. It remains to be seen whether it can make serious inroads into econometrics, but I firmly believe that it is a worthwhile investment for the younger cohorts of econometricians. R is public domain software and can be freely downloaded from the CRAN website. There is extensive documentation also available from CRAN under the heading manuals. For unix based systems it is usual to download R in source form, but it is also available in binary form for most common operating systems. There are several excellent introductions to R available in published form, in addition to the Introduction to R available in pdf from the CRAN website. I would particularly recommend Dalgaard (2002) and Venables and Ripley (2002). On the CRAN website there are also, under the heading "contributed", introductions to R in Danish, French, German, Spanish Italian, and a variety of other languages all of which can be freely downloaded in pdf. I've prepared a brief R FAQ that I will distribute with the course materials for the LSE short course.

For purposes of this course a minimal knowledge of R will suffice. R can be freely downloaded, and I hope that most students will bring a laptop so that they have access to R during the course sessions. Clicking the R icon should produce a window in which R will be running. To quit R, you just type `q()`, you will be prompted to answer whether you want to save the objects that were created during the session; responding "yes" will save the session objects into a file called .RData, responding

Version: May 11, 2011. These exercises were originally developed for a short course given under the auspices of CEMMAP at UCL, 20-22 February, 2003. My thanks to Andrew Chesher and the Department of Economics at UCL for their hospitality on that occasion, and as always to the NSF for continuing research support. The exercises have been expanded somewhat for new short courses under the auspices of CREATES in Aarhus, 21-23 June, 2010, and at the LSE, 16-17 May, 2011.

"no" will simply quit without saving. Online help is provided in two modes: if you know what you are looking for, you can type, for example `?rq` and you will get a description of the `rq` command, alternatively you can type `help.start()` and a browser help window should pop up and you can type more general key words or phrases to search for functionality.

R is intended to be a convenient interactive language and you can do many things on the fly by just typing commands into the R console, or even by pointing and clicking at one of the GUI interfaces, but I find that it is often preferable to save R commands into a file and execute a group of commands – this encourages a more reproducible style of research – and can be easily done using the `source("commands.R")` command. Saving output is a bit more complicated since there are many forms of output, graphics are usually saved in either postscript or pdf form, and tables can be saved in latex format for subsequent inclusion in documents. Together with Achim Zeileis, U. of Innsbruck, I've written a paper in *J. of Applied Econometrics* on reproducible research strategies that describes some of these things in more detail. The paper and some other ranting and raving about reproducibility are also available from my homepage by clicking first on "papers" and then on "Reproducible Econometric Research."

An aspect of reproducibility that is rarely considered in econometrics is the notion of "literate programming." The idea of literate programming was first broached by Donald Knuth in 1984; Knuth essentially advocated merging code and documentation for code in such a way that the code was self documenting and the exposition was self-documenting as well, since the code that generated the reported computations was embedded. In the R language this viewpoint has been implemented by Leisch's Sweave which can be considered to be a dialect of latex that allows the user to embed R chunks that are executed prior to latex compilation. The document that you are now reading was written in Sweave and can be viewed in its original form in the course problems directory as the file `ex.Rnw`.

Problem 1: A Family of Engel Curves

This is a simple bivariate linear quantile regression exercise designed to explore some basic features of the `quantreg` software in R. The data consists of observations on household food expenditure and household income of 235 working class Belgian familes taken from the well-known study of Ernst Engel (1857).

1. Read the data. The data can be downloaded from the website specified in class. You will see that it has a conventional ascii format with a header line indicating the variable names, and 235 lines of data, one per household. This can be read in R by the command

```
> url <- "http://www.econ.uiuc.edu/~roger/courses/LSE/data/engel.data"
> d <- read.table(file = url, header=TRUE) #data is now in matrix "d"
> attach(d) #attaching makes the variables accessible by name.
```

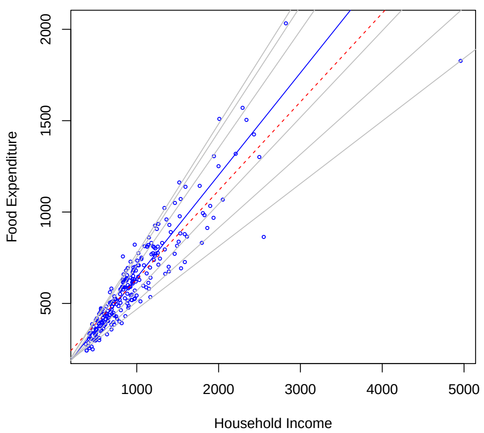
2. Plot the data. After the attach command the data is available using the names in the header, so we can plot the scatter diagram as:

```
> plot(x,y)
```

3. Replot with some better axis labels and superimpose some quantile regression lines on the scatter plot.

```
> require(quantreg)
> plot(x,y,cex=.25,type="n",xlab="Household Income",
         ylab="Food Expenditure")
> points(x,y,cex=.5,col="blue")
> abline(rq(y~x,tau=.5),col="blue")
```

```
> abline(lm(y~x),lty=2,col="red") #the dreaded ols line
> taus <- c(.05,.1,.25,.75,.90,.95)
> f <- rq(y ~ x, tau = taus)
> for( i in 1:length(taus)){
        abline(coef(f)[,i],col="gray")
        }
```



Note that you have to load the quantreg package before invoking the `rq()` command. Careful inspection of the plot reveals that the ols fit is severely biased at low incomes due to a few outliers. The plot command has a lot of options to fine tune the plot. There is a convenient looping structure, but beware that it can be slow in some applications. In `rq()` there are also many options: the first argument is a "formula" that specifies the model that is desired, in this case we want to fit the simple bivariate linear model so it is just `y~x` if we had two covariates we could say, e.g. `y~x+z`.

4. If we wanted to see all the distinct quantile regression solutions for this example we could specify a tau outside the range [0,1], e.g.

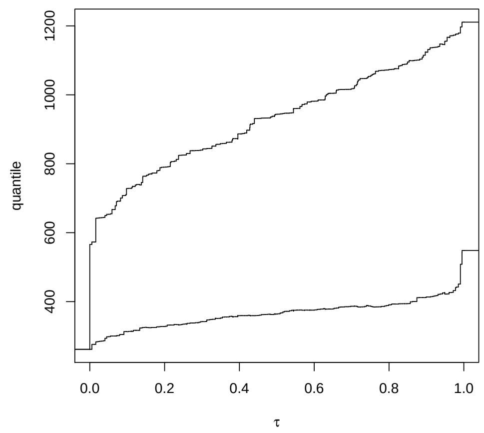
```
> z <- rq(y~x,tau=-1)
```

Now if you look at components of the structure `z` that are returned by the command, you can see for example the primal solution in `z$sol`, and the dual solution in `z$dsol`. In interactive mode just typing the name of some R object causes the program to print the object in some more or less easily intelligible manner. Now, if you want to estimate the conditional quantile function of `y` at a specific value of `x` and plot it you can do something like this:

```
> #Poor is defined as at the .1 quantile of the sample distn
> x.poor <- quantile(x,.1)
> #Rich is defined as at the .9 quantile of the sample distn
> x.rich <- quantile(x,.9)
```



```
> ps <- z$sol[1,]
> qs.poor <- c(c(1,x.poor)%*%z$sol[4:5,])
> qs.rich <- c(c(1,x.rich)%*%z$sol[4:5,])
> #now plot the two quantile functions to compare
> plot(c(ps,ps),c(qs.poor,qs.rich),type="n",
        xlab=expression(tau),ylab="quantile")
> plot(stepfun(ps,c(qs.poor[1],qs.poor)),do.points=FALSE,add=TRUE)
> plot(stepfun(ps,c(qs.poor[1],qs.rich)),do.points=FALSE,add=TRUE)
> #for conditional densities you could use akj()...
```



A nice feature of R is that documentation of functions usually includes some examples of their usage. These examples can be "run" by simply typing `example(SomeFunctionName)`, so for example when you type `example(rq)` you get a plot somewhat like the one you have just done "by hand." In a second plot you get a pair of coefficient plots that depict the estimate intercept and slope coefficients as a function of τ and provide a confidence band. More on this later. If you look carefully at the code being executing by in these examples you will see that you didn't need to download the data from the url specified, the Engel data is available directly from the **quantreg** package using the statement `data(engel)`. But it is often handy to be able to download data from the web. There are quite a lot of tools for handling web data sources, but this is another story entirely.

If you look carefully at the plots of the two estimated quantile functions that you made you will see minor violations of the expected monotonicity of these functions. This may or may not be regarded as a mortal sin, depending on your religious convictions. One way to deal with this, recently suggested by **?** is to "rearrange" the estimated functions. See the the documentation for this function using the usual strategy: `?rearrange`. To see an example of how this works try typing: `example(rearrange)`.

5. Now let's consider some formal testing. For starters suppose we just estimate two quartile fits and look at the default output:

```
> fit.25 <- rq(y~x,tau=.25)
> summary(fit.25)
Call: rq(formula = y ~ x, tau = 0.25)

tau: [1] 0.25

Coefficients:
            coefficients lower bd  upper bd
(Intercept)  95.48354     73.78608 120.09847
x             0.47410      0.42033   0.49433
> fit.75 <- rq(y~x,tau=.75)
> summary(fit.75)
Call: rq(formula = y ~ x, tau = 0.75)

tau: [1] 0.75

Coefficients:
            coefficients lower bd  upper bd
(Intercept)  62.39659     32.74488 107.31362
x             0.64401      0.58016   0.69041
```

By default the confidence intervals that are produced use the rank inversion method. This is fine for judging whether covariates are significant at particular quantiles but suppose that we wanted to test that the slopes were the same at the two quantiles? This is done with the `anova` command as follows:

```
> anova(fit.25,fit.75)
Quantile Regression Analysis of Deviance Table

Model: y ~ x
Joint Test of Equality of Slopes: tau in {  0.25 0.75  }

  Df Resid Df F value    Pr(>F)
1  1      469  30.891 4.586e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is an example of a general class of tests proposed in Koenker and Bassett (1982) It is instructive to look at the code for the command `anova.rq` to see how this test is carried out. The Wald approach is used and the asymptotic covariance matrix is estimated using the approach of Hendricks and Koenker (1991). It also illustrates a general syntax for testing in R adapted to the QR situation. If you have two models that are nested, with fits say `f0` and `f1`, then `anova(f0,f1)` should test whether the restricted model is correct. One needs to be careful however to check that the hypothesis that is intended, is really the one that the anova command understands, see `?anova.rq` for further details on the QR version of this. If you have more than two quantiles and want to do a joint test that all the slope coefficients are the same at all the quantiles you can use `anova(ft1,ft2,ft3,ft4)`.

In very large problems the rank inversion approach to confidence intervals is quite slow, and it is better to use another method. There are several choices. By default the computational method employs a variant of the Barrodale and Roberts (simplex-like) algorithm, for problems with sample size greater than about 5000 it is preferable to use interior point methods by using the `method="fn"`, flag in

the call to `rq`. When this "Frisch-Newton" version of the algorithm is used, rank test confidence intervals are not provided by summary instead a form of the Wald test is returned. Various options can be specified to produce various estimates of the standard errors as described below. These Wald forms of estimating standard errors are also possible to achieve with the default `method="br"` setting by adding for example the flag `se=nid`. Details of the algorithms are provided in Koenker and d'Orey (1987), Koenker and d'Orey (1993), for the "BR" method and Portnoy and Koenker (1997) for the "Frisch-Newton" method.

Standard inference results are obtained by calling summary, e.g.

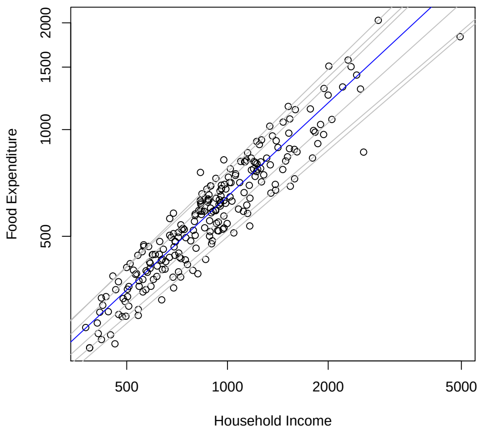
```
> fit <- rq(y~x,tau=.27,method="fn")
> summary(fit)
Call: rq(formula = y ~ x, tau = 0.27, method = "fn")

tau: [1] 0.27

Coefficients:
            coefficients lower bd  upper bd
(Intercept)  94.18652     81.53426 127.50707
x             0.48321      0.43213   0.50477
```

by default summary produces estimates of the asymptotic covariance matrix based on the approach described in Hendricks and Koenker (1991), an alternative approach suggested by Powell (1989) can be obtained by specifying `se="ker"`. There are further details and options regarding bandwidth and controlling the nature of what is returned by the summary command, see `?summary.rq` for these details.

At this point it would be useful to compare and contrast the various estimation and inference options that are available. Try estimating the simple model used above with both the `method = "br")` and `method = "fn")` choices, and then compare some of the `se` options in `summary.rq`.

6. The magic of logarithms. Thus far we have considered Engel functions that are linear in form, and the scatter as well as the QR testing has revealed a strong tendency for the dispersion of food expenditure to increase with household income. This is a particularly common form of heteroscedasticity. If one looks more carefully at the fitting, one sees interesting departures from symmetry that would not be likely to be revealed by the typical textbook testing for heteroscedasticity, however. One common remedy for symptoms like this would be to reformulate the model in log linear terms. It is interesting to compare what happens after the log transformation with what we have already seen. Consider the following plot:

```
> plot(x,y,log="xy",xlab="Household Income", ylab="Food Expenditure")
> taus <- c(.05,.1,.25,.75,.90,.95)
> abline(rq(log10(y)~log10(x),tau=.5),col="blue")
> for( i in 1:length(taus)){
        abline(rq(log10(y)~log10(x),tau=taus[i]),col="gray")
        }
```



Note that the flag `log="xy"` produces a plot with log-log axes, and for convenience of axis labeling these logarithms are base 10, so the subsequent fitting is also specified as base 10 logs for plotting purposes, even though base 10 logarithms are *unnatural* and would never be used in reporting numerical results. This looks much more like a classical iid error regression model, although again some departure from symmetry is visible. An interesting exercise is to conduct some formal testing for departures from the iid assumption of the type already considered above. This is left as an exercise for the reader.

Problem 2: Nonparametric Quantile Regression

Nonparametric quantile regression is most easily considered within a locally polynomial framework. Locally linear fitting can be carried out by the following function, provided in the `quantreg` package:

```
> "lprq" <-
 function(x, y, h, m=50 , tau=.5)
 {
         xx <- seq(min(x),max(x),length=m)
         fv <- xx
         der <- xx
         for(i in 1:length(xx)) {
                 z <- x - xx[i]
                 wx <- dnorm(z/h)
                 r <- rq(y~z, weights=wx, tau=tau)
                 fv[i] <- r$coef[1.]
                 der[i] <- r$coef[2.]
         }
         list(xx = xx, fv = fv, der = der)
 }
```


If you read through the function carefully you will see that it is just a matter of computing a quantile regression fit at each of m equally spaced x-values over the support of the observed x points. The function value estimates are returned as `fv` and the first derivative estimates at the m points are returned as `der`. As usual you can specify τ, but now you also need to specify a bandwidth `h`.

1. Begin by exploring the effect of the `h` and `tau` arguments for fitting the motorcycle data. Note that fitting derivatives requires larger bandwidth and larger sample size to achieve the same precision obtainable by function fitting. You are encouraged to substitute a more economic data set for the ubiquitous motorcycle data, its only advantage in the current context is that you can easily find examples to compare in the nonparametric regression literature.

2. Adapt `lprq` so that it does locally quadratic rather than linear fitting and compare performance.

3. Another general strategy for nonparametric quantile regression that is relatively simple to adapt to R uses regression splines. The function `bs()` in the package `splines` gives a very flexible way to construct B-spline basis expansions. For example you can fit a model like this:

```
> require(splines)
> url <- "http://www.econ.uiuc.edu/~roger/courses/LSE/data/motorcycle.data"
> d <- read.table(file = url, header=TRUE)
> fit <- rq(y~bs(x,df=5),tau=.33, data = d)
```

which fits a piecewise cubic polynomial with knots (breakpoints in the third derivative) at quintiles of the x's. You can also explicitly specify the knot sequence and the order of the spline. One advantage of this approach is that it is very easy to add a partially linear model component. So if there is another covariate, say z, we can add a parametric component like this:

```
> fit <- rq(y~bs(x,df=5)+z,tau=.33)
```

This avoids complications of backfitting when using kernel methods for partially linear models. Compare some fitting using the spline approach with that obtained with the local polynomial kernel approach.

4. Yet another even more appealing approach to univariate nonparametric smoothing involves penalty methods as described for example in Koenker, Ng, and Portnoy (1994) In recent work, Koenker and Mizera (2002), this approach has been extended to bivariate nonparametric regression, and more recently to a general class of additive models. Again, partially linear models are easily adapted, and there are easy ways to impose monotonicity and convexity on the fitted functions. In large problems it is essential to take advantage of the sparsity of the linear algebra. This is now feasible using special versions of the interior point algorithm for quantile regression and the SparseM package, Koenker and Ng (2003). The paper **?** describes some recent developments of inference apparatus for these models. Further development of these methods would be aided by some additional experience with real data.

An important feature of these additive models is that it is possible to impose monotonocity and/or convexity/concavity on the individual components. There are also relatively new methods for doing inference and prediction as well as plotting. As usual you can experiment with these methods by trying the `example()` function on methods like `summary.rqss`, `plot.rqss`, and `predict.rqss`. But more interesting would be to try new examples based on real data.

Problem 3: Quantile Regression Survival Analysis

Quantile regression as proven to be a particularly attractive approach for univariate survival analysis (aka duration modeling). The classical accelerated failure time model

$$\log(T_i) = x_i^\top \beta + u_i$$

with iid errors u_i, can be easily extended to consider,

$$Q_{\log(T_i)}(\tau|x_i) = x_i^\top \beta(\tau), \tag{1}$$

yielding a flexible, yet parametrically parsimonious, approach.

In this problem you are asked to explore such models in the context of the Pennsylvania reemployment bonus experiment conducted in 1988-89. In this period new claimants for unemployment insurance were randomized into one of several treatment groups or a control group. Control participants abided by the usual rules of the unemployment insurance system; treatment participants were offered a cash bonus to be awarded if the claimant was certifiably reemployed within a specified qualification period. For simplicity we will focus on only one of the treatment groups, those offered a bonus of 6 times their weekly benefit provided reemployment was established within 12 weeks. For this group the bonus averaged about $ 1000 for those collecting it. The data will be available in the form of an R data set called `Penn46.data` in the same directory as we have indicated for the prior datasets. This can be read into R using the same procedure as was used for the Engel data. For a more detailed analysis incorporating the other treatments, see Koenker and Bilias (2001). See Koenker and Xiao (2002) for further details on approaches to inference for these models.

In this application interest naturally focuses on the effect of the binary, randomized treatment. How does the bonus influence the distribution of the duration of unemployment? The Lehmann quantile treatment effect (QTE) is a natural object of empirical attention.

1. Explore some specifications of the QR model (1) and compare to fitting the Cox proportional hazard specification. See `require(survival)` for functions to estimate the corresponding Cox models. Note that covariate effects in the Cox models are necessarily scalar in nature, so for example the treatment effect must either increase, or decrease unemployment durations over the whole range of the distribution, but it cannot decrease durations in the lower tail and increase them in the upper tail – unless the model is specified with distinct baseline hazard functions for the two groups. See Koenker and Geling (2001) for some further details on the relationship between the QR survival model and the Cox model.
2. Explore some formal inference options to try to narrow the field of interesting specifications. See for example the discussion in Koenker and Xiao (2002) on tests based on the whole QR process.

Problem 4: Quantile Autoregression

Consider a simple linear QAR model,

$$y_t = \alpha_1(u_t) y_{t-1} + \alpha_0(u_t) \quad t = 0, 1, ..., T$$

where u_t is iid $U[0,1]$. Suppose that $\alpha_1(u) = 0.85 + 0.25u$ and $\alpha_0(u) = \Phi^{-1}(u)$ with Φ denoting the standard normal distribution function. Simulate a realization of this process with $T = 1000$ and estimate and plot the QAR coefficients, comparing them to the usual OLS estimates.

Verify whether or not the process is stationary. In your realization of the process check to see whether y_{t-1} stays in the region for which the conditional quantile function of y_t is monotone. What is the usual OLS estimate of the AR(1) model

estimating in this case? Check the residuals from the OLS fit to see if they exhibit any suspicious features that would reveal what is unusual here.

Problem 5: Portfolio Choice

This problem deals with the "pessimistic portfolio allocation" proposed in Bassett, Koenker, and Kordas (2003). The paper employs a highly artificial example. Your task, should you decide to accept it, is to produce a more realistic example using real data. Software implementing the methods of the paper is available as an R package called **`qrisk`**. This is *not* a CRAN package, but it is available from the url, `http://www.econ.uiuc.edu/~roger/research/risk/risk.html` The R function `qrisk` in this package computes optimal portfolio weights based on a matrix of observed, or simulated, asset returns using a specified form of pessimistic Choquet preferences.

Problem 6: Inequality Decomposition

The extensive literature on the measurement of inequality has devoted considerable attention to the question of how to decompose changes in measurements of inequality. If we observe increases in the Gini coefficient in a particular region over some sample period, can we attribute these changes in some way to underlying changes in covariates, or to changes in the effects of these covariates? QR offers a convenient general approach to this question. Suppose that we have estimated a garden variety wage equation model in QR form,

$$Q_{\log y}(\tau|x) = x^\top \beta(\tau), \tag{2}$$

and we would like to compute a conditional Gini coefficient.

Recall that the Lorenz function of a univariate distribution with quantile function, Q, is given by,

$$\lambda(t) = \mu^{-1} \int_0^t Q(s) ds$$

where $\mu = \int_0^1 Q(s) ds$ is the mean of the distribution. The Gini coefficient is simply twice the area between $\lambda(t)$ and the 45 degree line,

$$\gamma = 1 - 2 \int_0^1 \lambda(t) dt.$$

1. Given the linear decomposition of the conditional quantile function in (2) and the fact that the Gini coefficient is a linear functional of the quantile function, formulate a conditional Gini decomposition for log wages, and interpret it.
2. Over time we may wish to "explain" changes in *the* Gini coefficient by considering changes in the wage structure – which we can interpret as $\beta(\tau)$ in (2) – and changes in the characteristics of the population – which are captured by the evolution of the distribution of x. This way of thinking enables us to consider thought experiments such as, "How would Gini have evolved if the wage structure were fixed at some initial condition, but population characteristics changed according to some specified pattern, historical or otherwise". Or alternatively, suppose that we fix population characteristics and consider the evolution of the the conditional components of Gini as $\beta_t(\tau)$ changes over time. Decompositions of this type have been considered in recent work of Machado and Mata (2001). The Gini decomposition has also been recently considered by **?** I would love to see a further applications along these lines.

References

Bassett, G., R. Koenker, and G. Kordas (2003): "Pessimistic Portfolio Allocation and Choquet Expected Utility," preprint.

Dalgaard, P. (2002): *Introductory Statistics with R*. Springer.

Hendricks, W., and R. Koenker (1991): "Hierarchical spline models for conditional quantiles and the demand for elec tricity," *J. of Am. Stat. Assoc.*, 87, 58–68.

Koenker, R., and G. Bassett (1982): "Robust tests for heteroscedasticity based on regression quantiles," *Econometrica*, 50, 43–61.

Koenker, R., and Y. Bilias (2001): "Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments," *Empirical Economics*, 26, 199–220.

Koenker, R., and V. d'Orey (1987): "Computing Regression Quantiles," *Applied Statistics*, 36, 383–393.

——— (1993): "A Remark on Computing Regression Quantiles," *Applied Statistics*, 36, 383–393.

Koenker, R., and O. Geling (2001): "Reappraising Medfly Longevity: A quantile regression survival analysis," *J. of Am. Stat. Assoc.*, 96, 458–468.

Koenker, R., and I. Mizera (2002): "Penalized Triograms: Total Variation Regularization for Bivariate Smoothing," preprint.

Koenker, R., and P. Ng (2003): "SparseM: A Sparse Linear Algebra Package for R," preprint.

Koenker, R., P. Ng, and S. Portnoy (1994): "Quantile Smoothing Splines," *Biometrika*, 81, 673–80.

Koenker, R., and Z. Xiao (2002): "Inference on the quantile regression process," *Econometrica*, 70, 1583–1612.

Machado, J., and J. Mata (2001): "Counterfactual decomposition of changes in wage distributions using quantile regression," *Empirical Economics*, 26, 115–134.

Portnoy, S., and R. Koenker (1997): "The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators, with discusssion," *Stat. Science*, 12, 279–300.

Powell, J. L. (1989): "Estimation of monotonic regression models under quantile restrictions," in *Nonparametric and Semiparametric Methods in Econometrics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge U. Press: Cambridge.

Venables, W., and B. Ripley (2002): *Modern Applied Statistics with S*. Springer.

Quantile Regression: A Gentle Introduction

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 16 May 2011

Overview of the Course

- The Basics: What, Why and How?
- Inference and Quantile Treatment Effects
- Nonparametric Quantile Regression
- Endogoneity and IV Methods
- Censored QR and Survival Analysis
- Quantile Autoregression
- QR for Longitudinal Data
- Risk Assessment and Choquet Portfolios
- Computional Aspects

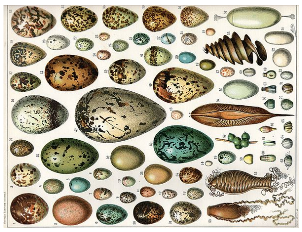
Course outline, lecture slides, an R FAQ, and even some proposed exercises can all be found at:

`http://www.econ.uiuc.edu/~roger/courses/LSE`.

The Basics: What, Why and How?

1. Univariate Quantiles
2. Scatterplot Smoothing
3. Equivariance Properties
4. Quantile Treatment Effects
5. Three Empirical Examples

Archimedes' "Eureka!" and the Middle Sized Egg



Volume of the eggs can be measure by the amount of water they displace and the median (middle-sized) egg found by sorting these measurements.

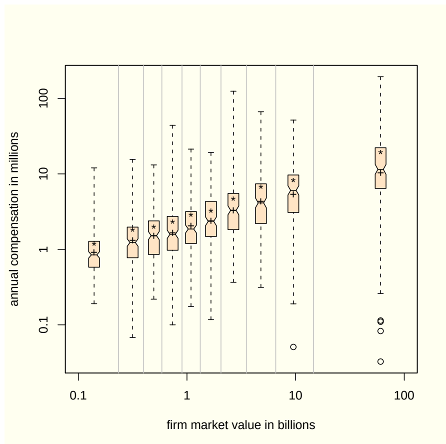
Note that even if we measure the logarithm of the volumes, the middle sized egg is the same! Not true for the mean egg, or the modal one.

The Stem and Leaf Plot: Tukey's EDA Gadget Number 1

Given a "batch" of numbers, $\{X_1, X_2, \ldots, X_n\}$ one can make a quick and dirty histogram in R this way:

```
> x <- rchisq(100,5) # 100 Chi-squared(5)
> quantile(x) # Tukey's Five Number Summary
       0%        25%        50%        75%       100%
 0.9042396  2.7662230  4.2948642  6.2867588 16.5818573

> stem(x)

  The decimal point is at the |

   0 | 92356668
   2 | 001111244445667778889990111222455666
   4 | 0122333466667890112556788 9
   6 | 023344667802888
   8 | 556691
  10 | 7
  12 | 159
  14 | 06
  16 | 6
```


Boxplot of CEO Pay: Tukey's EDA Gadget Number 2



Motivation

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

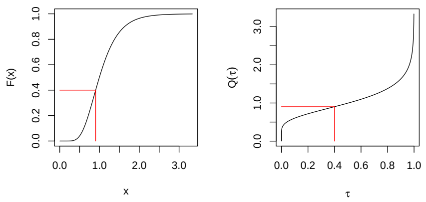
Mosteller and Tukey (1977)

Univariate Quantiles

Given a real-valued random variable, X, with distribution function F, we will define the τth quantile of X as

$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x \mid F(x) \geqslant \tau\}.$$

This definition follows the usual convention that F is CADLAG, and Q is CAGLAD as illustrated in the following pair of pictures.

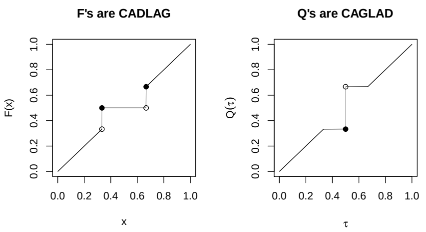


Univariate Quantiles

Given a real-valued random variable, X, with distribution function F, we will define the τth quantile of X as

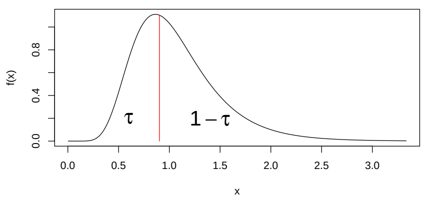
$$Q_X(\tau) = F_X^{-1}(\tau) = \inf\{x \mid F(x) \geqslant \tau\}.$$

This definition follows the usual convention that F is CADLAG, and Q is CAGLAD as illustrated in the following pair of pictures.



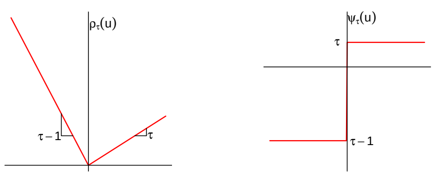
Univariate Quantiles

Viewed from the perspective of densities, the τth quantile splits the area under the density into two parts: one with area τ below the τth quantile and the other with area $1 - \tau$ above it:



Two Bits Worth of Convex Analysis

A convex function ρ and its subgradient ψ:



The subgradient of a convex function $f(u)$ at a point u consists of all the possible "tangents." Sums of convex functions are convex.

Population Quantiles as Optimizers

Quantiles solve a simple optimization problem:

$$\hat{\alpha}(\tau) = \operatorname{argmin} \mathbb{E}\, \rho_\tau(Y - \alpha)$$

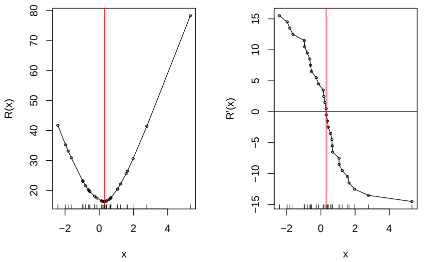
Proof: Let $\psi_\tau(u) = \rho'_\tau(u)$, so differentiating wrt to α:

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \psi_\tau(y - \alpha)dF(y) \\ &= (\tau - 1)\int_{-\infty}^{\alpha} dF(y) + \tau \int_{\alpha}^{\infty} dF(y) \\ &= (\tau - 1)F(\alpha) + \tau(1 - F(\alpha)) \end{aligned}$$

implying $\tau = F(\alpha)$ and thus $\hat{\alpha} = F^{-1}(\tau)$.

Sample Quantiles as Optimizers

For sample quantiles replace F by $\hat{F}$, the empirical distribution function. The objective function becomes a polyhedral convex function whose derivative is monotone decreasing, in effect the gradient simply counts observations above and below and weights the sums by τ and $1-\tau$.



Conditional Quantiles: The Least Squares Meta-Model

The unconditional mean solves

$$\mu = \text{argmin}_m \mathbb{E}(Y-m)^2$$

The conditional mean $\mu(x) = E(Y|X=x)$ solves

$$\mu(x) = \text{argmin}_m \mathbb{E}_{Y|X=x}(Y-m(X))^2.$$

Similarly, the unconditional τth quantile solves

$$\alpha_\tau = \text{argmin}_a \mathbb{E}\rho_\tau(Y-a)$$

and the conditional τth quantile solves

$$\alpha_\tau(x) = \text{argmin}_a \mathbb{E}_{Y|X=x}\rho_\tau(Y-a(X))$$

Computation of Linear Regression Quantiles

Primal Formulation as a linear program, split the residual vector into positive and negative parts and sum with appropriate weights:

$$\min\{\tau 1^\top u + (1-\tau)1^\top v | y = Xb + u - v, (b,u,v) \in \mathbb{R}^p \times \mathbb{R}^{2n}_+\}$$

Dual Formulation as a Linear Program

$$\max\{y'd | X^\top d = (1-\tau)X^\top 1, d \in [0,1]^n\}$$

Solutions are characterized by an exact fit to p observations.
Let $h \in \mathcal{H}$ index p-element subsets of $\{1,2,\ldots,n\}$ then primal solutions take the form:

$$\hat{\beta} = \hat{\beta}(h) = X(h)^{-1}y(h)$$

Least Squares from the Quantile Regression Perspective

Exact fits to p observations:

$$\hat{\beta} = \hat{\beta}(h) = X(h)^{-1}y(h)$$

OLS is a weighted average of these $\hat{\beta}(h)$'s:

$$\hat{\beta}_{OLS} = (X^\top X)^{-1}X^\top y = \sum_{h\in\mathcal{H}} w(h)\hat{\beta}(h),$$

$$w(h) = |X(h)|^2 / \sum_{h\in\mathcal{H}} |X(h)|^2$$

The determinants $|X(h)|$ are the (signed) volumes of the parallelipipeds formed by the columns of the the matrices $X(h)$. In the simplest bivariate case, we have,

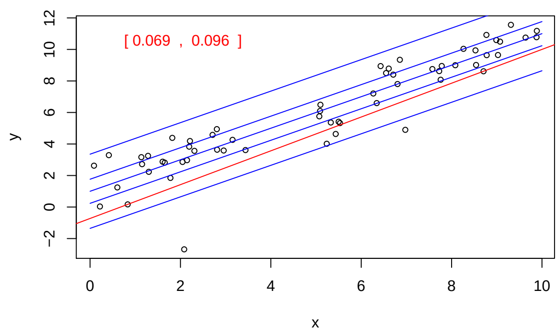
$$|X(h)|^2 = \begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix}^2 = (x_j - x_i)^2$$

so pairs of observations that are far apart are given more weight.

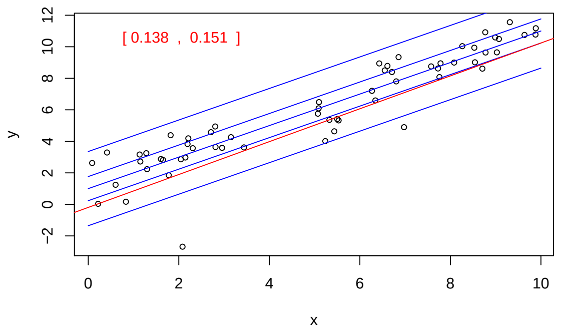
Quantile Regression: The Movie

- Bivariate linear model with iid Student t errors
- Conditional quantile functions are parallel in blue
- 100 observations indicated in blue
- Fitted quantile regression lines in red.
- Intervals for $\tau \in (0,1)$ for which the solution is optimal.

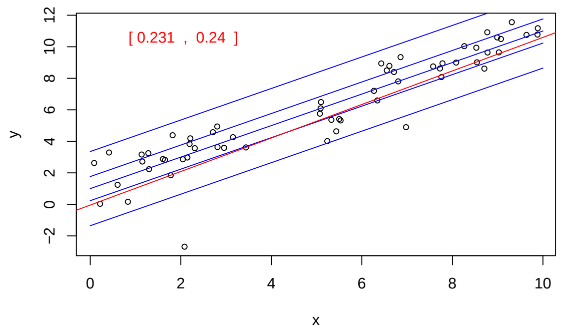
Quantile Regression in the iid Error Model



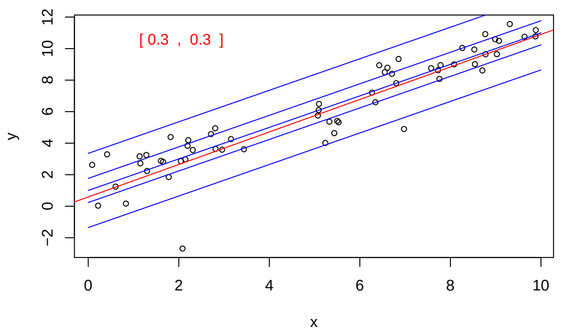
Quantile Regression in the iid Error Model



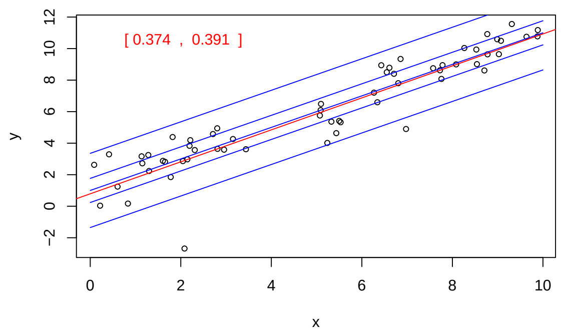
Quantile Regression in the iid Error Model



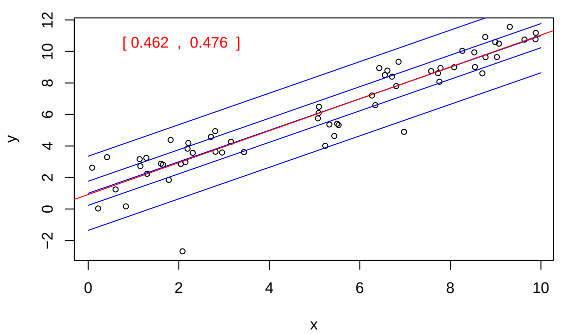
Quantile Regression in the iid Error Model



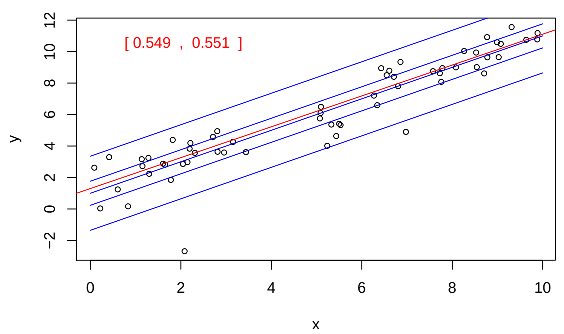
Quantile Regression in the iid Error Model



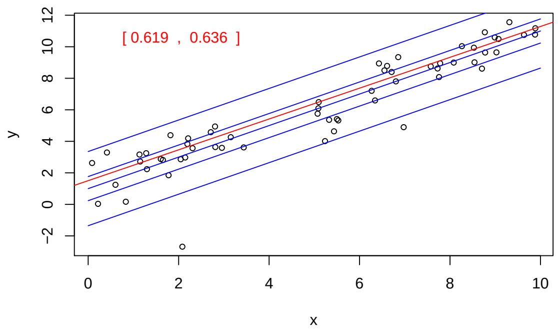
Quantile Regression in the iid Error Model



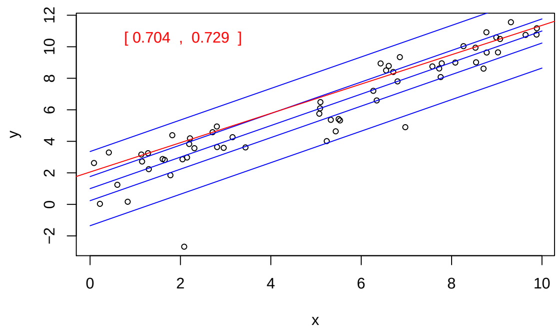
Quantile Regression in the iid Error Model



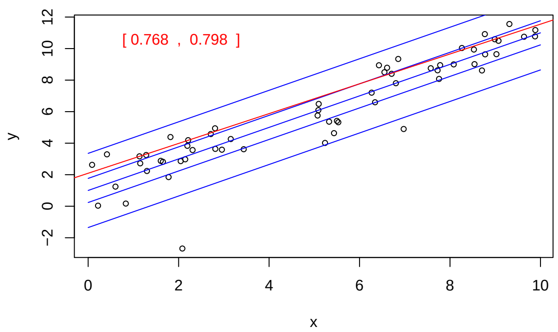
Quantile Regression in the iid Error Model



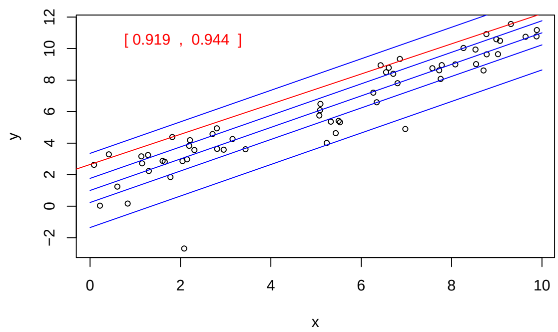
Quantile Regression in the iid Error Model



Quantile Regression in the iid Error Model



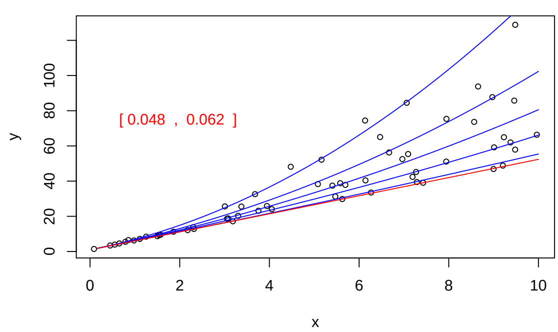
Quantile Regression in the iid Error Model



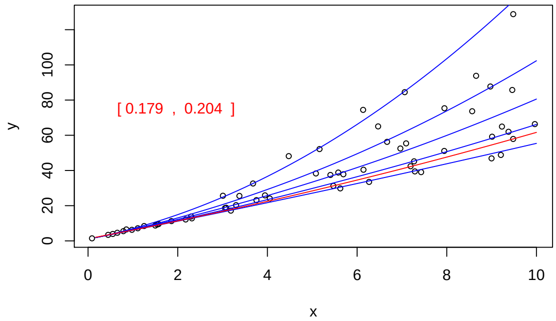
Virtual Quantile Regression II

- Bivariate quadratic model with Heteroscedastic χ^2 errors
- Conditional quantile functions drawn in blue
- 100 observations indicated in blue
- Fitted quadratic quantile regression lines in red
- Intervals of optimality for $\tau \in (0, 1)$.

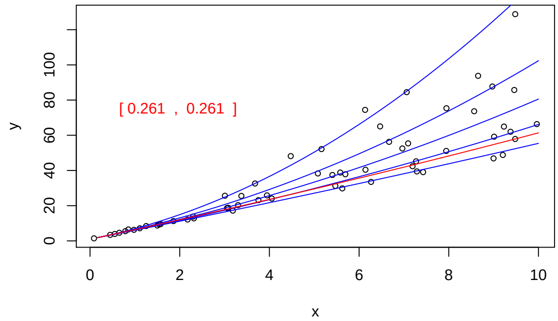
Quantile Regression in the Heteroscedastic Error Model



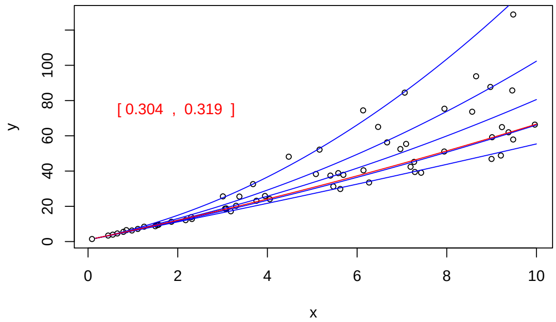
Quantile Regression in the Heteroscedastic Error Model



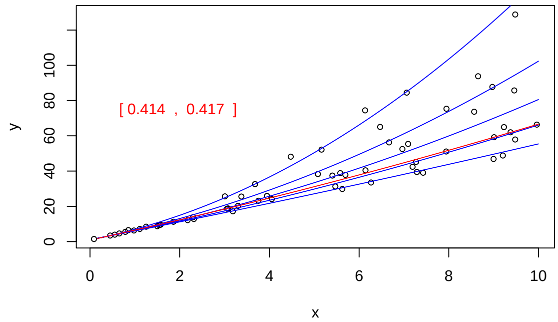
Quantile Regression in the Heteroscedastic Error Model



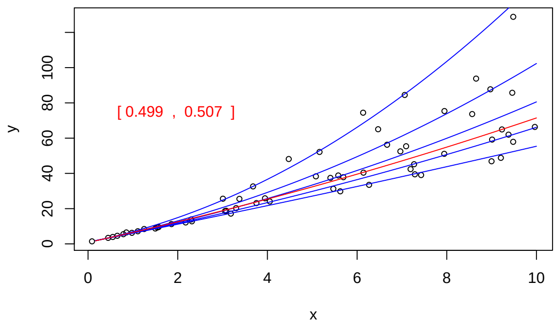
Quantile Regression in the Heteroscedastic Error Model



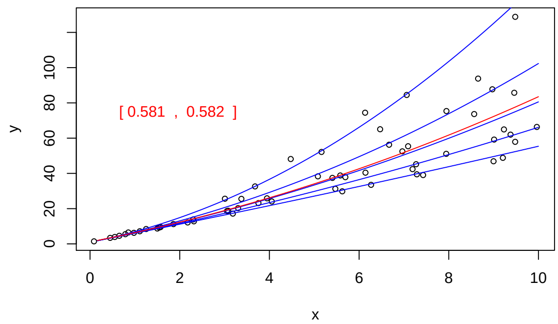
Quantile Regression in the Heteroscedastic Error Model



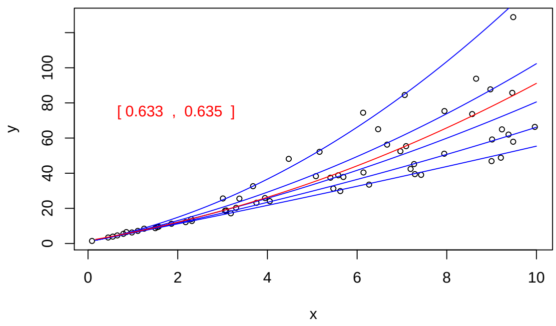
Quantile Regression in the Heteroscedastic Error Model



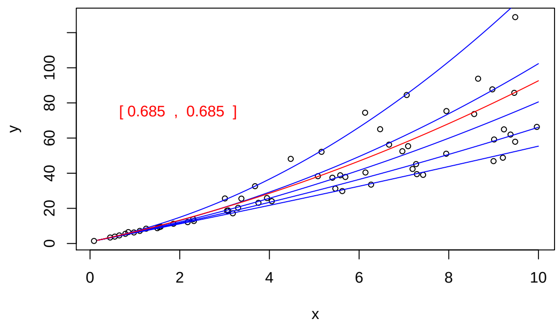
Quantile Regression in the Heteroscedastic Error Model



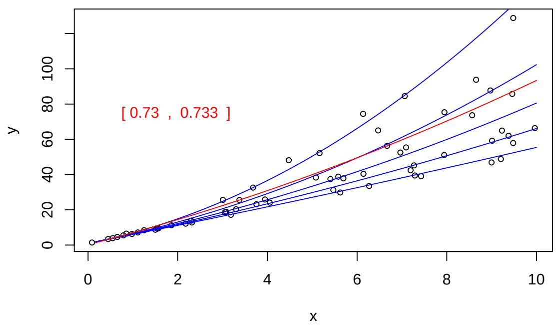
Quantile Regression in the Heteroscedastic Error Model



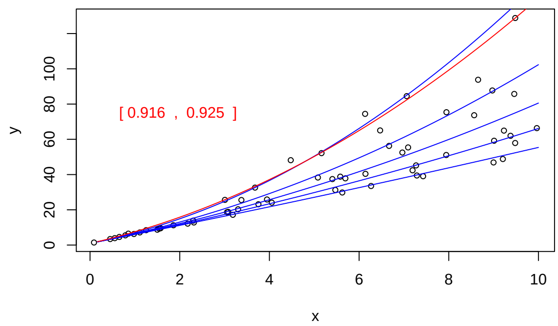
Quantile Regression in the Heteroscedastic Error Model



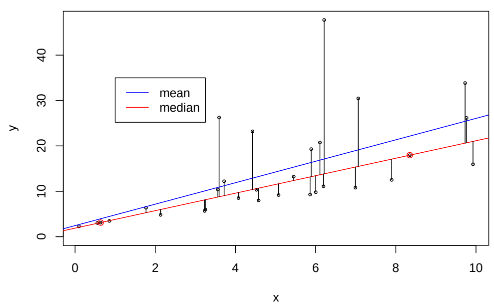
Quantile Regression in the Heteroscedastic Error Model



Quantile Regression in the Heteroscedastic Error Model



Conditional Means vs. Medians



Minimizing absolute errors for median regression can yield something quite different from the least squares fit for mean regression.

Equivariance of Regression Quantiles

- Scale Equivariance: For any $a > 0$, $\hat{\beta}(\tau; ay, X) = a\hat{\beta}(\tau; y, X)$ and $\hat{\beta}(\tau; -ay, X) = a\hat{\beta}(1-\tau; y, X)$
- Regression Shift: For any $\gamma \in \mathbb{R}^p$ $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y, X) + \gamma$
- Reparameterization of Design: For any $|A| \neq 0$, $\hat{\beta}(\tau; y, AX) = A^{-1}\hat{\beta}(\tau; yX)$
- Robustness: For any diagonal matrix D with nonnegative elements. $\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau, y + D\hat{u}, X)$

Equivariance to Monotone Transformations

For any monotone function h, conditional quantile functions $Q_Y(\tau|x)$ are equivariant in the sense that

$$Q_{h(Y)|X}(\tau|x) = h(Q_{Y|X}(\tau|x))$$

In contrast to conditional mean functions for which, generally,

$$E(h(Y)|X) \neq h(EY|X)$$

Examples:
$h(y) = \min\{0, y\}$, Powell's (1985) censored regression estimator.
$h(y) = \mathrm{sgn}\{y\}$ Rosenblatt's (1957) perceptron, Manski's (1975) maximum score estimator. estimator.

Beyond Average Treatment Effects

Lehmann (1974) proposed the following general model of treatment response:

> *"Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F."*

Lehmann QTE as a QQ-Plot

Doksum (1974) defines $\Delta(x)$ as the "horizontal distance" between F and G at x, *i.e.*

$$F(x) = G(x + \Delta(x)).$$

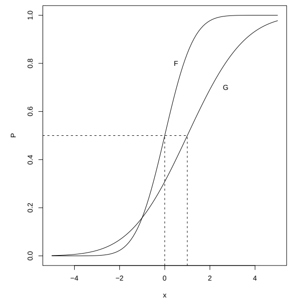
Then $\Delta(x)$ is uniquely defined as

$$\Delta(x) = G^{-1}(F(x)) - x.$$

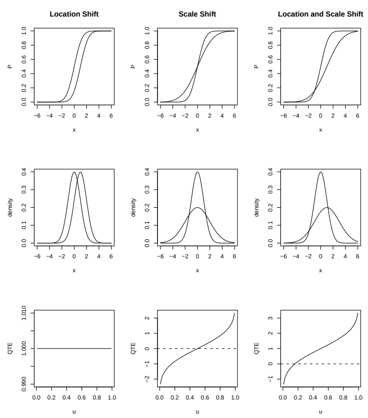
This is the essence of the conventional QQ-plot. Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$

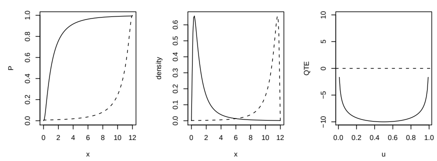
Lehmann-Doksum QTE



Lehmann-Doksum QTE



An Asymmetric Example



Treatment shifts the distribution from right skewed to left skewed making the QTE U-shaped.

The Erotic is Unidentified

The Lehmann QTE characterizes the difference in the marginal distributions, F and G, but it cannot reveal anything about the joint distribution, H. The copula function, Schweizer and Wolf (1981), Genest and McKay, (1986),

$$\varphi(u, v) = H(F^{-1}(u), G^{-1}(v)),$$

is *not* identified. Lehmann's formulation *assumes* that the treatment leaves the ranks of subjects invariant. If a subject was going to be the median control subject, then he will also be the median treatment subject. This is an inherent limitation of the Neymann-Rubin potential outcomes framework.

QTE via Quantile Regression

The Lehmann QTE is naturally estimable by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau)$$

where $\hat{G}_n$ and $\hat{F}_m$ denote the empirical distribution functions of the treatment and control observations, Consider the quantile regression model

$$Q_{Y_i}(\tau | D_i) = \alpha(\tau) + \delta(\tau) D_i$$

where D_i denotes the treatment indicator, and $Y_i = h(T_i)$, *e.g.* $Y_i = \log T_i$, which can be estimated by solving,

$$\min \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i)$$

Francis Galton's (1885) Anthropometric Quantiles

224 *NATURE* [*Jan.* 8, 1885

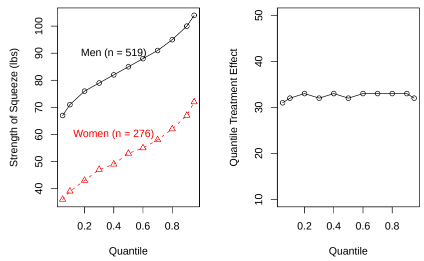
ANTHROPOMETRIC PER-CENTILES

Values surpassed, and Values unreached, by various percentages of the persons measured at the Anthropometric Laboratory in the late International Health Exhibition

(*The value that is unreached by n per cent. of any large group of measurements, and surpassed by 100–n of them, is called its nth percentile*)

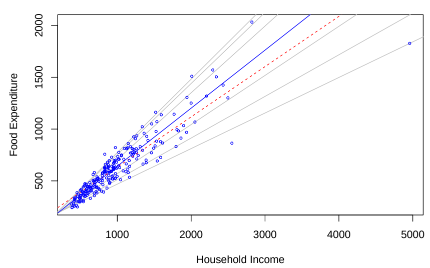
Subject of measurement	Age	Unit of measurement	Sex	No. of persons in the group	Values surpassed by per-cents as below: 95 / Values unreached by per-cents. as below: 5	90 / 10	80 / 20	70 / 30	60 / 40	50 / 50	40 / 60	30 / 70	20 / 80	10 / 90	5 / 95
Height, standing, without shoes ...	23–51	Inches	M.	811	63·2	64·5	65·8	66·5	67·3	67·9	68·5	69·2	70·0	71·3	72·4
			F.	770	58·8	59·9	61·3	62·1	62·7	63·3	63·9	64·6	65·3	66·4	67·3
Height, sitting, from seat of chair ...	23–51	Inches	M.	1013	33·6	34·2	34·9	35·3	35·4	36·0	36·3	36·7	37·1	37·7	38·2
			F.	775	31·8	32·3	32·9	33·3	33·6	33·9	34·2	34·6	34·9	35·6	36·0
Span of arms ...	23–51	Inches	M.	811	65·0	66·1	67·2	68·2	69·0	69·9	70·6	71·4	72·3	73·6	74·8
			F.	770	58·6	59·5	60·7	61·7	62·4	63·0	63·7	64·5	65·4	66·7	68·0
Weight in ordinary indoor clothes ...	23–26	Pounds	M.	520	121	125	131	135	139	143	147	150	156	165	172
			F.	276	102	105	110	114	118	122	129	132	136	142	149
Breathing capacity	23–26	Cubic inches	M.	212	161	177	187	199	211	219	226	236	248	277	290
			F.	277	92	102	115	124	131	138	144	151	164	177	186
Strength of pull as archer with bow	23–26	Pounds	M.	519	56	60	64	68	71	74	77	88	82	89	96
			F.	276	30	32	34	36	38	40	42	44	47	51	54
Strength of squeeze with strongest hand	23–26	Pounds	M.	519	67	71	76	79	82	85	88	91	95	100	104
			F.	276	36	39	43	47	49	52	55	58	62	67	72
Swiftness of blow.	23–26	Feet per second	M.	516	13·2	14·1	15·2	16·2	17·3	18·1	19·1	20·0	20·9	22·3	23·6
			F.	271	9·2	10·1	11·3	12·1	12·8	13·4	14·0	14·5	15·1	16·3	16·9
Sight, keenness of—by distance of reading diamond test-type	23–26	Inches	M.	398	13	17	20	22	23	25	26	28	30	32	34
			F.	433	10	12	16	19	22	24	26	27	29	31	32

Quantile Treatment Effects: Strength of Squeeze



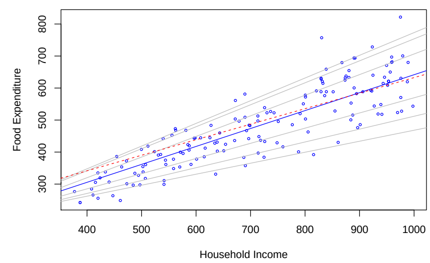
"Very powerful women exist, but happily perhaps for the repose of the other sex, such gifted women are rare."

Engel's Food Expenditure Data



Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the blue solid line; the least squares estimate of the conditional mean function is indicated by the red dashed line.

Engel's Food Expenditure Data

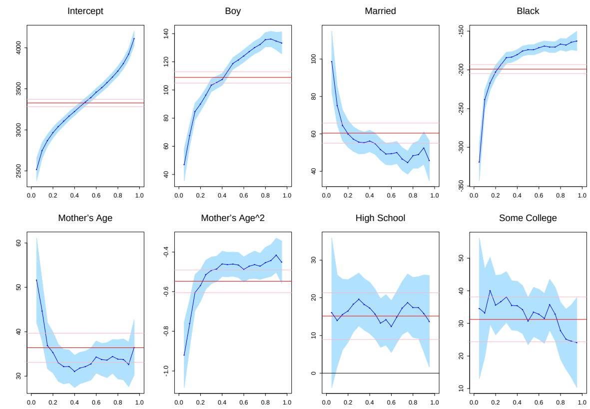


Engel Curves for Food: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for $\tau \in \{.05, .1, .25, .5, .75, .9, .95\}$ are superimposed on the scatterplot. The median $\tau = .5$ fit is indicated by the blue solid line; the least squares estimate of the conditional mean function is indicated by the red dashed line.

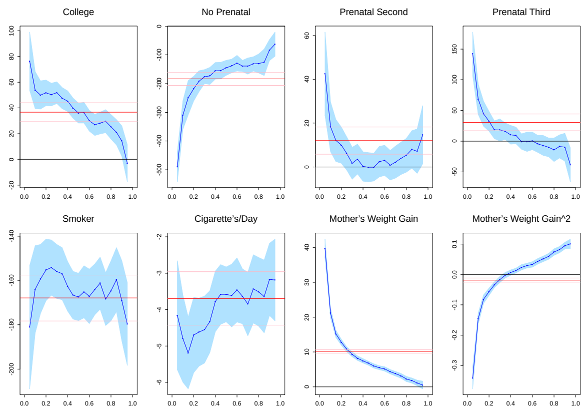
A Model of Infant Birthweight

- Reference: Abrevaya (2001), Koenker and Hallock (2001)
- Data: June, 1997, Detailed Natality Data of the US. Live, singleton births, with mothers recorded as either black or white, between 18-45, and residing in the U.S. Sample size: 198,377.
- Response: Infant Birthweight (in grams)
- Covariates:
  - Mother's Education
  - Mother's Prenatal Care
  - Mother's Smoking
  - Mother's Age
  - Mother's Weight Gain

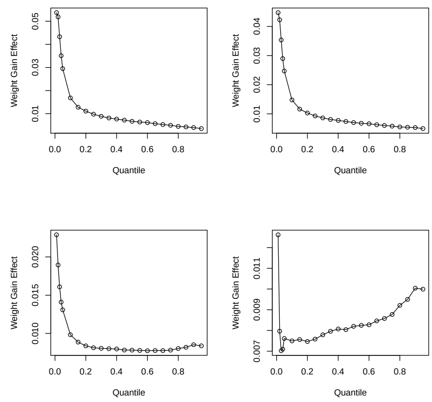
Quantile Regression Birthweight Model I



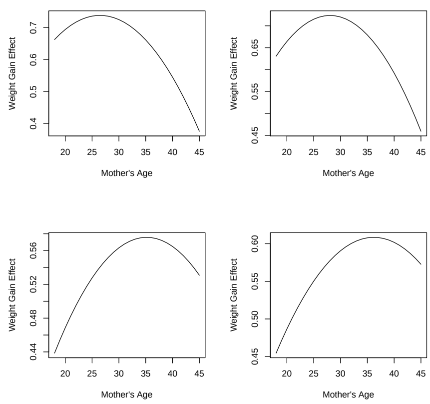
Quantile Regression Birthweight Model II



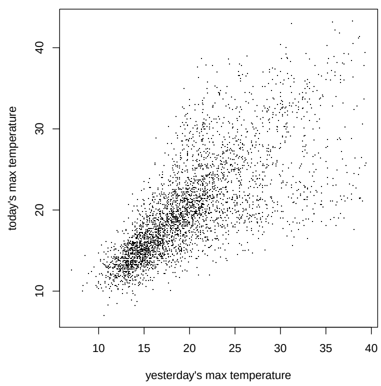
Marginal Effect of Mother's Age



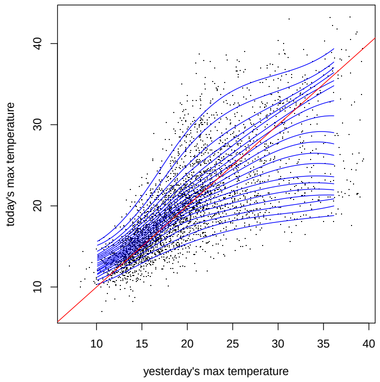
Marginal Effect of Mother's Weight Gain



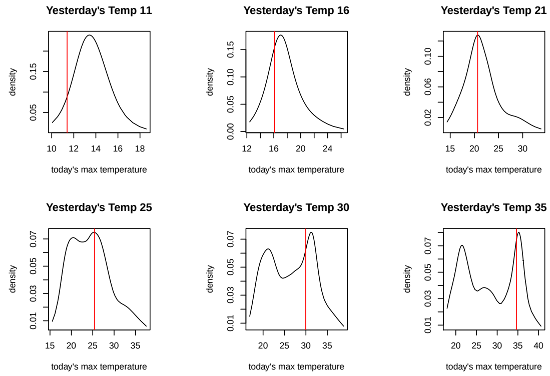
Daily Temperature in Melbourne: AR(1) Scatterplot



Daily Temperature in Melbourne: Nonlinear QAR(1) Fit



Conditional Densities of Melbourne Daily Temperature



Review of Lecture 1

Least squares meethods of estimating conditional mean functions

- were developed for, and
- promote the view that,

$$\text{Response} = \text{Signal} + \text{iid Measurement Error}$$

In fact the world is rarely this simple.

Quantile Regression: Inference

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 16 May 2011

Inference for Quantile Regression

- Inference for the Sample Quantiles
- QR Inference in iid Error Models*
- QR Inference in Heteroscedastic Error Models*
- Classical Rank Tests and the Quantile Regression Dual*
- Inference on the Quantile Regression Process*

* Skimmed very lightly in favor of the first DIY in R session.

What determines the precision of sample quantiles?

For random samples from a continuous distribution, F, the sample quantiles, $\hat{F}_n^{-1}(\tau)$ are consistent, by the Glivenko-Cantelli theorem. Rates of convergence and precision are governed by the density near the quantile of interest, if it exists.
Note that differentiating the identity: $F(F^{-1}(t)) = t$, yields,

$$\frac{d}{dt}F(F^{-1}(t)) = f(F^{-1}(t))\frac{d}{dt}F^{-1}(t) = 1$$

thus, provided of course that $f(F^{-1}(t)) > 0$,

$$\frac{d}{dt}F^{-1}(t) = \frac{1}{f(F^{-1}(t))}$$

So, limiting normality of $\hat{F}_n$ and the δ-method imply limiting normality of the sample quantiles with $\sqrt{n}$ rate and variance proportional to $f^{-2}(F^{-1}(t))$.

Inference for the Sample Quantiles

Minimizing $\sum_{i=1}^n \rho_\tau(y_i - \xi)$ consider

$$g_n(\xi) = -n^{-1}\sum_{i=1}^n \psi_\tau(y_i - \xi) = n^{-1}\sum_{i=1}^n (I(y_i < \xi) - \tau).$$

By convexity of the objective function,

$$\{\hat{\xi}_\tau > \xi\} \Leftrightarrow \{g_n(\xi) < 0\}$$

and the DeMoivre-Laplace CLT yields, expanding F,

$$\sqrt{n}(\hat{\xi}_\tau - \xi) \rightsquigarrow \mathcal{N}(0, \omega^2(\tau, F))$$

where $\omega^2(\tau, F) = \tau(1-\tau)/f^2(F^{-1}(\tau))$. Classical Bahadur-Kiefer representation theory provides further refinement of this result.

Some Gory Details

Instead of a fixed $\xi = F^{-1}(\tau)$ consider,

$$\mathbb{P}\{\hat{\xi}_n > \xi + \delta/\sqrt{n}\} = \mathbb{P}\{g_n(\xi + \delta/\sqrt{n}) < 0\}$$

where $g_n \equiv g_n(\xi + \delta/\sqrt{n})$ is a sum of iid terms with

$$\begin{aligned}
\mathbb{E}g_n &= \mathbb{E}n^{-1}\sum_{i=1}^n (I(y_i < \xi + \delta/\sqrt{n}) - \tau) \\
&= F(\xi + \delta/\sqrt{n}) - \tau \\
&= f(\xi)\delta/\sqrt{n} + o(n^{-1/2}) \\
&\equiv \mu_n\delta + o(n^{-1/2})
\end{aligned}$$

$$\mathbb{V}g_n = \tau(1-\tau)/n + o(n^{-1}) \equiv \sigma_n^2 + o(n^{-1}).$$

Thus, by (a triangular array form of) the DeMoivre-Laplace CLT,

$$\mathbb{P}(\sqrt{n}(\hat{\xi}_n - \xi) > \delta) = \Phi((0 - \mu_n\delta)/\sigma_n) \equiv 1 - \Phi(\omega^{-1}\delta)$$

where $\omega = \mu_n/\sigma_n = \sqrt{\tau(1-\tau)}/f(F^{-1}(\tau))$.

Finite Sample Theory for Quantile Regression

Let $h \in \mathcal{H}$ index the $\binom{n}{p}$ p-element subsets of $\{1, 2, \dots, n\}$ and $X(h), y(h)$ denote corresponding submatrices and vectors of X and y.
Lemma: $\hat{\beta} = b(h) \equiv X(h)^{-1}y(h)$ is the τth regression quantile iff $\xi_h \in \mathcal{C}$ where

$$\xi_h = \sum_{i \notin h} \psi_\tau(y_i - x_i\hat{\beta})x_i^\top X(h)^{-1},$$

$\mathcal{C} = [\tau - 1, \tau]^p$, and $\psi_\tau(u) = \tau - I(u < 0)$.
Theorem: (KB, 1978) In the linear model with iid errors, $\{u_i\} \sim F, f$, the density of $\hat{\beta}(\tau)$ is given by

$$\begin{aligned}
g(b) = \; & \textstyle\sum_{h \in \mathcal{H}} \prod_{i \in h} f(x_i^\top(b - \beta(\tau)) + F^{-1}(\tau)) \\
& \cdot P(\xi_h(b) \in C)|\det(X(h))|
\end{aligned}$$

Asymptotic behavior of $\hat{\beta}(\tau)$ follows by (painful) consideration of the limiting form of this density, see also Knight and Goh (ET, 2009).

Asymptotic Theory of Quantile Regression I

In the classical linear model,

$$y_i = x_i \beta + u_i$$

with u_i iid from dfF, with density $f(u) > 0$ on its support $\{u | 0 < F(u) < 1\}$, the joint distribution of $\sqrt{n}(\hat{\beta}_n(\tau_i) - \beta(\tau_i))_{i=1}^m$ is asymptotically normal with mean 0 and covariance matrix $\Omega \otimes D^{-1}$. Here $\beta(\tau) = \beta + F_u^{-1}(\tau) e_1, e_1 = (1, 0, \ldots, 0)^\top$, $x_{1i} \equiv 1, n^{-1} \sum x_i x_i^\top \to D$, a positive definite matrix, and

$$\Omega = ((\tau_i \wedge \tau_j - \tau_i \tau_j)/(f(F^{-1}(\tau_i)) f(F^{-1}(\tau_j)))_{i,j=1}^m.$$

Asymptotic Theory of Quantile Regression II

When the response is conditionally independent over i, but not identically distributed, the asymptotic covariance matrix of $\zeta(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ is somewhat more complicated. Let $\xi_i(\tau) = x_i \beta(\tau)$, $f_i(\cdot)$ denote the corresponding conditional density, and define,

$$\begin{aligned} J_n(\tau_1, \tau_2) &= (\tau_1 \wedge \tau_2 - \tau_1 \tau_2) n^{-1} \sum_{i=1}^n x_i x_i^\top, \\ H_n(\tau) &= n^{-1} \sum x_i x_i^\top f_i(\xi_i(\tau)). \end{aligned}$$

Under mild regularity conditions on the $\{f_i\}$'s and $\{x_i\}$'s, we have joint asymptotic normality for $(\zeta(\tau_i), \ldots, \zeta(\tau_m))$ with covariance matrix

$$V_n = (H_n(\tau_i)^{-1} J_n(\tau_i, \tau_j) H_n(\tau_j)^{-1})_{i,j=1}^m.$$

Making Sandwiches

The crucial ingredient of the QR Sandwich is the quantile density function $f_i(\xi_i(\tau))$, which can be estimated by a difference quotient. Differentiating the identity: $F(Q(t)) = t$ we get

$$s(t) = \frac{dQ(t)}{dt} = \frac{1}{f(Q(t))}$$

sometimes called the "sparsity function" so we can compute

$$\hat{f}_i(x_i^\top \hat{\beta}(\tau)) = 2h_n / (x_i^\top (\hat{\beta}(\tau + h_n) - \hat{\beta}(\tau - h_n))$$

with $h_n = O(n^{-1/3})$. Prudence suggests a modified version:

$$\tilde{f}_i(x_i^\top \hat{\beta}(\tau)) = \max\{0, \hat{f}_i(x_i^\top \hat{\beta}(\tau))\}$$

Various other strategies can be employed including a variety of bootstrapping options. More on this in the first lab session.

Rank Based Inference for Quantile Regression

- Ranks play a fundamental *dual* role in QR inference.
- Classical rank tests for the p-sample problem extended to regression
- Rank tests play the role of Rao (score) tests for QR.

Two Sample Location-Shift Model

$$\begin{aligned} &X_1, \ldots, X_n \sim F(x) && \text{“Controls”} \\ &Y_1, \ldots, Y_m \sim F(x - \theta) && \text{“Treatments”} \end{aligned}$$

Hypothesis:

$$\begin{aligned} H_0 &: \quad \theta = 0 \\ H_1 &: \quad \theta > 0 \end{aligned}$$

The Gaussian Model $F = \Phi$

$$T = (\bar{Y}_m - \bar{X}_n) / \sqrt{n^{-1} + m^{-1}}$$

UMP Tests:

$$\text{critical region } \{T > \Phi^{-1}(1 - \alpha)\}$$

Wilcoxon-Mann-Whitney Rank Test

Mann-Whitney Form:

$$S = \sum_{i=1}^n \sum_{j=1}^m I(Y_j > X_i)$$

Heuristic: If treatment responses are larger than controls for most pairs (i, j), then H_0 should be rejected.
Wilcoxon Form: Set $(R_1, \ldots, R_{n+m}) = \text{Rank}(Y_1, \ldots, Y_m, X_1, \ldots X_n)$,

$$W = \sum_{j=1}^m R_j$$

Proposition: $S = W - m(m+1)/2$ so Wilcoxon and Mann-Whitney tests are equivalent.

Pros and Cons of the Transformation to Ranks

Thought One:
Gain: Null Distribution is independent of F.
Loss: Cardinal information about data.

Thought Two:
Gain: Student t-test has quite accurate size provided $\sigma^2(F) < \infty$.
Loss: Student t-test uses cardinal information badly for long-tailed F.

Asymptotic Relative Efficiency of Wilcoxon versus Student t-test

Pitman (Local) Alternatives: $H_n : \theta_n = \theta_0/\sqrt{n}$
$(\text{t-test})^2 \rightsquigarrow \chi_1^2(\theta_0^2/\sigma^2(F))$
$(\text{Wilcoxon})^2 \rightsquigarrow \chi_1^2(12\theta_0^2(\int f^2)^2)$

$$\text{ARE}(W, t, F) = 12\sigma^2(F)[\int f^2(x)dx]^2$$

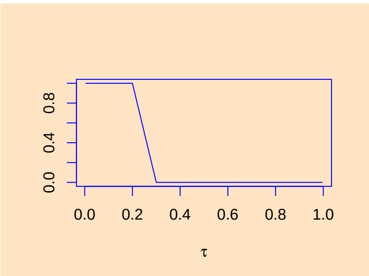
F	N	U	Logistic	DExp	LogN	t_2
ARE	.955	1.0	1.1	1.5	7.35	∞

Theorem (Hodges-Lehmann) For all F, $ARE(W, t, F) \geqslant .864$.

Hájek 's Rankscore Generating Functions

Let $Y_1, \ldots, Y_n$ be a random sample from an absolutely continuous df F with associated ranks $R_1, \ldots, R_n$, Hájek 's rank generating functions are:

$$\hat{a}_i(t) = \begin{cases} 1 & \text{if } t \leqslant (R_i - 1)/n \\ R_i - tn & \text{if } (R_i - 1)/n \leqslant t \leqslant R_i/n \\ 0 & \text{if } R_i/n \leqslant t \end{cases}$$



Linear Rank Statistics Asymptotics

Theorem (Hájek (1965)) Let $c_n = (c_{1n}, \ldots, c_{nn})$ be a triangular array of real numbers such that

$$\max_i (c_{in} - \bar{c}_n)^2 / \sum_{i=1}^n (c_{in} - \bar{c}_n)^2 \to 0.$$

Then

$$\begin{aligned} Z_n(t) &= (\sum_{i=1}^n (c_{in} - \bar{c}_n)^2)^{-1/2} \sum_{j=1}^n (c_{jn} - \bar{c}_n)\hat{a}_j(t) \\ &\equiv \sum_{j=1}^n w_j \hat{a}_j(t) \end{aligned}$$

converges weakly to a Brownian Bridge, i.e., a Gaussian process on $[0, 1]$ with mean zero and covariance function $\text{Cov}(Z(s), Z(t)) = s \wedge t - st$.

Some Asymptotic Heuristics

The Hájek functions are approximately indicator functions

$$\hat{a}_i(t) \approx I(Y_i > F^{-1}(t)) = I(F(Y_i) > t)$$

Since $F(Y_i) \sim U[0, 1]$, linear rank statistics may be represented as

$$\int_0^1 \hat{a}_i(t)d\varphi(t) \approx \int_0^1 I(F(Y_i) > t)d\varphi(t) = \varphi(F(Y_i)) - \varphi(0)$$

$$\begin{aligned} \int_0^1 Z_n(t)d\varphi(t) &= \sum w_i \int \hat{a}_i(t)d\varphi(t) \\ &= \sum w_i \varphi(F(Y_i)) + o_p(1), \end{aligned}$$

which is asymptotically distribution free, i.e. independent of F.

Duality of Ranks and Quantiles

Quantiles may be *defined* as

$$\hat{\xi}(\tau) = \text{argmin} \sum \rho_\tau(y_i - \xi)$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$. This can be formulated as a linear program whose dual solution

$$\hat{a}(\tau) = \text{argmax}\{y^\top a | 1_n^\top a = (1 - \tau)n, a \in [0, 1]^n\}$$

generates the Hájek rankscore functions.

Reference: Gutenbrunner and Jurečková (1992).

Regression Quantiles and Rank Scores:

$$\hat{\beta}_n(\tau) = \text{argmin}_{b \in \mathbb{R}^p} \sum \rho_\tau(y_i - x_i^\top b)$$

$$\hat{a}_n(\tau) = \text{argmax}_{a \in [0,1]^n} \{y^\top a | X^\top a = (1-\tau) X^\top 1_n\}$$

$x^\top \hat{\beta}_n(\tau)$ — Estimates $Q_Y(\tau|x)$
Piecewise constant on $[0,1]$.
For $X = 1_n$, $\hat{\beta}_n(\tau) = \hat{F}_n^{-1}(\tau)$.

$\{\hat{a}_i(\tau)\}_{i=1}^n$ — Regression rankscore functions
Piecewise linear on $[0,1]$.
For $X = 1_n$, $\hat{a}_i(\tau)$ are Hajek rank generating functions.

Regression Rankscore "Residuals"

The Wilcoxon rankscores,

$$\tilde{u}_i = \int_0^1 \hat{a}_i(t) dt$$

play the role of quantile regression residuals. For each observation y_i they answer the question: on which quantile does y_i lie? The $\tilde{u}_i$ satisfy an orthogonality restriction:

$$X^\top \tilde{u} = X^\top \int_0^1 \hat{a}(t) dt = n\bar{x} \int_0^1 (1-t) dt = n\bar{x}/2.$$

This is something like the $X^\top \hat{u} = 0$ condition for OLS. Note that if the X is "centered" then $\bar{x} = (1, 0, \cdots, 0)$. The $\tilde{u}$ vector is approximately uniformly "distributed;" in the one-sample setting $u_i = (R_i + 1/2)/n$ so they are obviously "too uniform."

Regression Rank Tests

$$Y = X\beta + Z\gamma + u$$

$$H_0 : \gamma = 0 \text{ versus } H_n : \gamma = \gamma_0/\sqrt{n}$$

Given the regression rank score process for the restricted model,

$$\hat{a}_n(\tau) = \text{argmax}\left\{Y^\top a \,|\, X^\top a = (1-\tau) X^\top 1_n\right\}$$

A test of H_0 is based on the linear rank statistics,

$$\hat{b}_n = \int_0^1 \hat{a}_n(t)\, d\varphi(t)$$

Choice of the score function φ permits test of location, scale or (potentially) other effects.

Regression Rankscore Tests

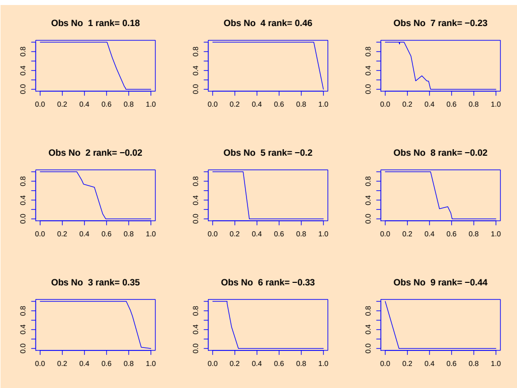
Theorem: (Gutenbrunner, Jurečková , Koenker and Portnoy) Under H_n and regularity conditions, the test statistic $T_n = S_n^\top Q_n^{-1} S_n$ where $S_n = (Z - \hat{Z})^\top \hat{b}_n$, $\hat{Z} = X(X^\top X)^{-1} X^\top Z$, $Q_n = n^{-1}(Z - \hat{Z})^\top Z - \hat{Z})$

$$T_n \rightsquigarrow \chi_q^2(\eta)$$

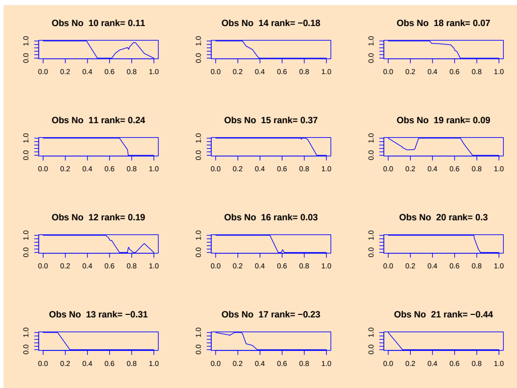
where

$$\eta^2 = \omega^2(\varphi, F)\gamma_0^\top Q \gamma_0$$
$$\omega(\varphi, F) = \int_0^1 f(F^{-1}(t))\, d\varphi(t)$$

Regression Rankscores for Stackloss Data



Regression Rankscores for Stackloss Data



Inversion of Rank Tests for Confidence Intervals

For the scalar γ case and using the score function

$$\varphi_\tau(t) = \tau - I(t < \tau)$$

$$\hat{b}_{ni} = -\int_0^1 \varphi_\tau(t) d\hat{a}_{ni}(t) = \hat{a}_{ni}(\tau) - (1-\tau)$$

where $\bar{\varphi} = \int_0^1 \varphi_\tau(t)dt = 0$ and $A^2(\varphi_\tau) = \int_0^1 (\varphi_\tau(t) - \bar{\varphi})^2 dt = \tau(1-\tau)$.
Thus, a test of the hypothesis $H_0 : \gamma = \xi$ may be based on $\hat{a}_n$ from solving,

$$\max\{(y - x_2\xi)^\top a | X_1^\top a = (1-\tau)X_1^\top 1, a \in [0,1]^n\} \quad (1)$$

and the fact that

$$S_n(\xi) = n^{-1/2} x_2^\top \hat{b}_n(\xi) \rightsquigarrow \mathcal{N}(0, A^2(\varphi_\tau) q_n^2) \quad (2)$$

Inversion of Rank Tests for Confidence Intervals

That is, we may compute

$$T_n(\xi) = S_n(\xi)/(A(\varphi_\tau) q_n)$$

where $q_n^2 = n^{-1} x_2^\top (I - X_1(X_1^\top X_1)^{-1} X_1^\top) x_2$. and reject H_0 if $|T_n(\xi)| > \Phi^{-1}(1-\alpha/2)$.

Inverting this test, that is finding the interval of ξ's such that the test fails to reject. This is a quite straightforward parametric linear programming problem and provides a simple and effective way to do inference on individual quantile regression coefficients. Unlike the Wald type inference it delivers asymmetric intervals. This is the default approach to parametric inference in `quantreg` for problems of modest sample size.

Inference on the Quantile Regression Process

Using the quantile score function, $\varphi_\tau(t) = \tau - I(t < \tau)$ we can consider the quantile rankscore process,

$$T_n(\tau) = S_n(\tau)^\top Q_n^{-1} S_n(\tau)/(\tau(1-\tau)).$$

where

$$\begin{aligned} S_n &= n^{-1/2}(X_2 - \hat{X}_2)^\top \hat{b}_n, \\ \hat{X}_2 &= X_1(X_1^\top X_1)^{-1} X_1^\top X_2, \\ Q_n &= (X_2 - \hat{X}_2)^\top (X_2 - \hat{X}_2)/n, \\ \hat{b}_n &= (-\int \varphi(t) d\hat{a}_{in}(t))_{i=1}^n, \end{aligned}$$

Inference on the Quantile Regression Process

Theorem: (K & Machado) Under $H_n : \gamma(\tau) = O(1/\sqrt{n})$ for $\tau \in (0,1)$ the process $T_n(\tau)$ converges to a non-central Bessel process of order $q = \dim(\gamma)$. Pointwise, T_n is non-central χ^2.

Related Wald and LR statistics can be viewed as providing a general apparatus for testing goodness of fit for quantile regression models. This approach is closely related to classical p-dimensional goodness of fit tests introduced by Kiefer (1959).
When the null hypotheses under consideration involve unknown nuisance parameters things become more interesting. In Koenker and Xiao (2001) we consider this "Durbin problem" and show that the elegant approach of Khmaladze (1981) yields practical methods.

Four Concluding Comments about Inference

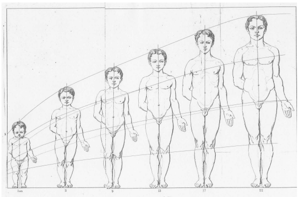
- Asymptotic inference for quantile regression poses some statistical challenges since it involves elements of nonparametric density estimation, but this shouldn't be viewed as a major obstacle.
- Classical rank statistics and Hájek 's rankscore process are closely linked via Gutenbrunner and Jurečková 's regression rankscore process, providing an attractive approach to many inference problems while avoiding density estimation.
- Inference on the quantile regression process can be conducted with the aid of Khmaladze's extension of the Doob-Meyer construction.
- Resampling offers many further lines of development for inference in the quantile regression setting.

Nonparametric Quantile Regression

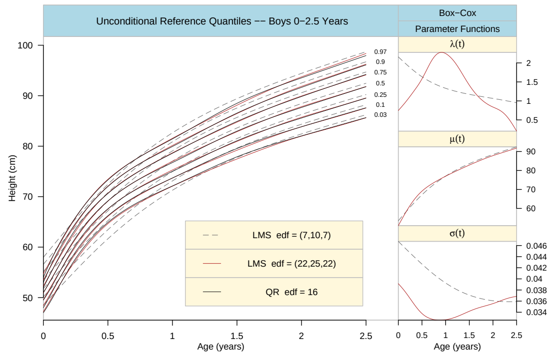
Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011



In the Beginning, . . . were the Quantiles



Pere, Wei, He, and K *Stat. in Medicine* (2006)

Three Approaches to Nonparametric Quantile Regression

- Locally Polynomial (Kernel) Methods: `lprq`

$$\hat{\alpha}(\tau, x) = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau\left(y_i - \alpha_0 - \alpha_1(x_i - x) - \ldots - \frac{1}{p!}\alpha_p(x_i - x)^p\right)$$
$$\hat{g}(\tau, x) = \hat{\alpha}_0(\tau, x)$$

- Series Methods `rq( y ~ bs(x,knots = k) + z`

$$\hat{\alpha}(\tau) = \operatorname{argmin}_\alpha \sum_{i=1}^{n} \rho_\tau\left(y_i - \sum_j \varphi_j(x_i)\alpha_j\right)$$
$$\hat{g}(\tau, x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\alpha}_j$$

- Penalty Methods `rqss`

$$\hat{g}(\tau, x) = \operatorname{argmin}_g \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda P(g)$$

Total Variation Regularization I

There are many possible penalties, ways to measure the roughness of fitted function, but total variation of the first derivative of g is particularly attractive:

$$P(g) = V(g') = \int |g''(x)|dx$$

As $\lambda \to \infty$ we constrain g to be closer to linear in x. Solutions of

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda V(g')$$

are continuous and piecewise linear.

Example 1: Fish in a Bottle

Objective: to study metabolic activity of various fish species in an effort to better understand the nature of the feeding cycle. Metabolic rates based on oxygen consumption as measured by sensors mounted on the tubes.

Three primary aspects are of interest:

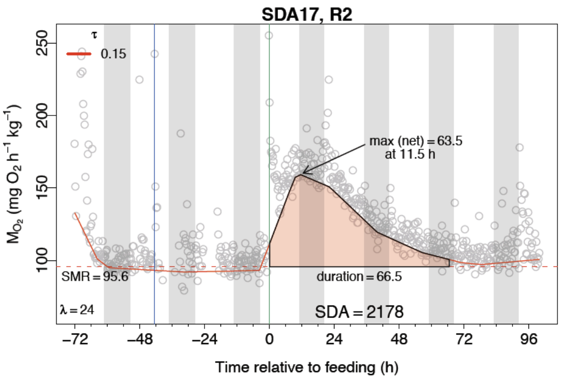
1. Basal (minimal) Metabolic Rate,
2. Duration and Shape of the Feeding Cycle, and
3. Diurnal Cycle.

Example 1: Some Experimental Details

Experimental data of Denis Chabot, Institut Maurice-Lamontagne, Quebec, Canada and his colleagues.

1. Basal (minimal) metabolic rate M_{O_2} (aka Standard Metabolic Rate SMR) is measured in $mg\ O_2\ h^{-1}\ kg^{-1}$ for fish "at rest" after several days without feeding,
2. Fish are then fed and oxygen consumption monitored until M_{O_2} returns to its prior SMR level for several hours.
3. Elevation of M_{O_2} after feeding (aka Specific Dynamic Action SDA) ideally measures the energy required for digestion,
4. Procedure is repeated for several cycles, so each estimation of the cycle is based on a few hundred observations.

Example 1: Juvenile Codfish



Tuning Parameter Selection

There are two tuning parameters:

1. $\tau = 0.15$ the (low) quantile chosen to represent the SMR,
2. λ controls the smoothness of the SDA cycle.

One way to interpret the parameter λ is to note that it controls the number of effective parameters of the fitted model (Meyer and Woodroofe(2000):

$$p(\lambda) = \text{div } \hat{g}_{\lambda,\tau}(y_1, ..., y_n) = \sum_{i=1}^{n} \partial \hat{y}_i / \partial y_i$$

This is equivalent to the number of interpolated observations, the number of zero residuals. Selection of λ can be made by minimizing, e.g. Schwarz Criterion:

$$\text{SIC}(\lambda) = n \log(n^{-1} \sum \rho_\tau(y_i - \hat{g}_{\lambda,\tau}(x_i))) + \frac{1}{2} p(\lambda) \log n.$$

Total Variation Regularization II

For bivariate functions we consider the analogous problem:

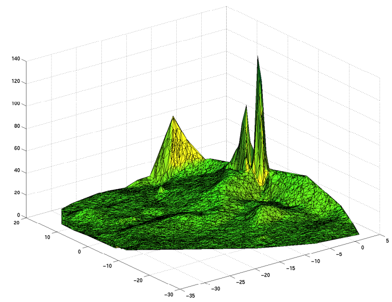
$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_{1i}, x_{2i})) + \lambda V(\nabla g)$$

where the total variation variation penalty is now:

$$V(\nabla g) = \int \|\nabla^2 g(x)\| dx$$

Solutions are again continuous, but now they are piecewise linear on a triangulation of the observed x observations. Again, as $\lambda \to \infty$ solutions are forced toward linearity.

Example 2: Chicago Land Values via TV Regularization



Chicago Land Values: Based on 1194 vacant land sales and 7505 "virtual" sales introduced to increase the flexibility of the triangulation. K and Mizera (2004).

Additive Models: Putting the pieces together

We can combine such models:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - \sum_j g_j(x_{ij})) + \sum_j \lambda_j V(\nabla g_j)$$

- Components g_j can be univariate, or bivariate.
- Additivity is intended to muffle the curse of dimensionality.
- Linear terms are easily allowed, or enforced.
- And shape restrictions like monotonicity and convexity/concavity as well as boundry conditions on g_j's can also be imposed.

Implementation in the R `quantreg` Package

- Problems are typically large, very sparse linear programs.
- Optimization via interior point methods are quite efficient,
- Provided sparsity of the linear algebra is exploited, quite large problems can be estimated.
- The nonparametric `qss` components can be either univariate, or bivariate
- Each `qss` component has its own λ specified
- Linear covariate terms enter formula in the usual way
- The `qss` components can be shape constrained.

```
fit <- rqss(y ~ qss(x1,3) + qss(x2,8) + x3, tau = .6)
```

Pointwise Confidence Bands

It is obviously crucial to have reliable confidence bands for nonparametric components. Following Wahba (1983) and Nychka(1983), conditioning on the λ selection, we can construct bands from the covariance matrix of the full model:

$$V = \tau(1-\tau)(\tilde{X}^\top \Psi \tilde{X})^{-1}(\tilde{X}^\top \tilde{X})^{-1}(\tilde{X}^\top \Psi \tilde{X})^{-1}$$

with

$$\tilde{X} = \begin{bmatrix} X & G_1 & \cdots & G_J \\ \lambda_0 H_K & 0 & \cdots & 0 \\ 0 & \lambda_1 P_1 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_J P_J \end{bmatrix} \quad \text{and} \quad \Psi = \mathrm{diag}(\phi(\hat{u}_i/h_n)/h_n)$$

Pointwise bands can be constructed by extracting diagonal blocks of V.

Uniform Confidence Bands

Uniform bands are also important, but more challenging. We would like:

$$B_n(x) = (\hat{g}_n(x) - c_\alpha \hat{\sigma}_n(x), \hat{g}_n(x) + c_\alpha \hat{\sigma}_n(x))$$

such that the true curve, g_0, is covered with specified probability $1-\alpha$ over a given domain $\mathcal{X}$:

$$\mathcal{P}\{g_0(x) \in B_n(x) \mid x \in \mathcal{X}\} \geqslant 1 - \alpha.$$

We can follow the "Hotelling tube" approach based on Hotelling(1939) and Weyl (1939) as developed by Naiman (1986), Johansen and Johnstone (1990) Sun and Loader (1994) and others.

Uniform Confidence Bands

Hotelling's original formulation for parametric nonlinear regression has been extended to non-parametric regression. For series estimators

$$\hat{g}_n(x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\theta}_j$$

with pointwise standard error $\sigma(x) = \sqrt{\varphi(x)^\top V^{-1} \varphi(x)}$ we would like to invert test statistics of the form:

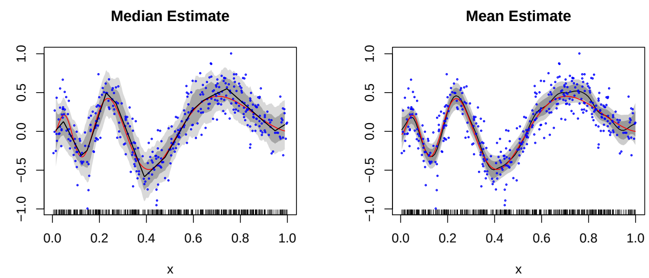
$$T_n = \sup_{x \in \mathcal{X}} \frac{\hat{g}_n(x) - g_0(x)}{\sigma(x)}.$$

This requires solving for the critical value, c_α in

$$\mathcal{P}(T_n > c) \leqslant \frac{\kappa}{2\pi}(1 + c^2/\nu)^{-\nu/2} + \mathcal{P}(t_\nu > c) = \alpha$$

where κ is the length of a "tube" determined by the basis expansion, t_ν is a Student random variable with degrees of freedom $\nu = n - p$.

Confidence Bands in Simulations



$$Y_i = \sqrt{x_i(1-x_i)} \sin\left(\frac{2\pi(1+2^{-7/5})}{x_i + 2^{-7/5}}\right) + U_i, \quad i = 1, \cdots, 400, \quad U_i \sim \mathcal{N}(0, 0.04)$$

,

Simulation Performance

	Accuracy			Pointwise		Uniform	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.063	0.046	12.936	0.960	0.999	0.323	0.920
gam	0.045	0.035	20.461	0.956	0.998	0.205	0.898
t_3							
rqss	0.071	0.052	11.379	0.955	0.998	0.274	0.929
gam	0.071	0.054	17.118	0.948	0.994	0.159	0.795
t_1							
rqss	0.099	0.070	9.004	0.930	0.996	0.161	0.867
gam	35.551	2.035	8.391	0.920	0.926	0.203	0.546
χ^2_3							
rqss	0.110	0.083	8.898	0.950	0.997	0.270	0.883
gam	0.096	0.074	14.760	0.947	0.987	0.218	0.683

Performance of Penalized Estimators and Their Confidence Bands: IID Error Model

Simulation Performance

	Accuracy			Pointwise		Uniform	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.081	0.063	10.685	0.951	0.998	0.265	0.936
gam	0.064	0.050	17.905	0.957	0.999	0.234	0.940
t_3							
rqss	0.091	0.070	9.612	0.952	0.998	0.241	0.938
gam	0.103	0.078	14.656	0.949	0.992	0.232	0.804
t_1							
rqss	0.122	0.091	7.896	0.938	0.997	0.222	0.893
gam	78.693	4.459	7.801	0.927	0.958	0.251	0.695
χ^2_3							
rqss	0.145	0.114	7.593	0.947	0.998	0.307	0.921
gam	0.138	0.108	12.401	0.941	0.973	0.221	0.626

Performance of Penalized Estimators and Their Confidence Bands: Linear Scale Model

Example 3: Childhood Malnutrition in India

A larger scale problem illustrating the use of these methods is a model of risk factors for childhood malnutrition considered by Fenske, Kneib and Hothorn (2009).

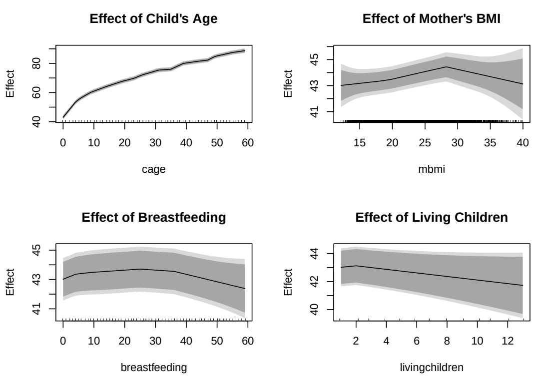
- They motivate the use of models for low conditional quantiles of height as a way to explore influences on malnutrition,
- They employ boosting as a model selection device,
- Their model includes six univariate nonparametric components and 15 other linear covariates.
- There are 37,623 observations on the height of children from India.

Example 3: R Formulation

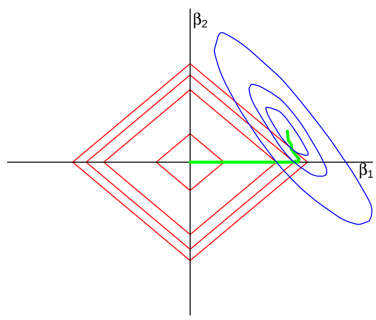
```
 fit <- rqss(cheight ~ qss(cage, lambda = lam[1]) +
qss(bfed, lambda = lam[2]) + qss(mage, lambda = lam[3]) +
qss(mbmi, lambda = lam[4]) + qss(sibs, lambda = lam[5]) +
qss(medu, lambda = lam[6]) + qss(fedu, lambda = lam[7]) +
csex + ctwin + cbirthorder + munemployed + mreligion +
mresidence + deadchildren + wealth + electricity +
radio + television + frig + bicycle + motorcycle + car +
tau = 0.10, method = "lasso", lambda = lambda, data = india)
```

- The seven coordinates of `lam` control the smoothness of the nonparametric components,
- `lambda` controls the degree of shrinkage in the linear (lasso) coefficients.
- The estimated model has roughly 40,000 observations, including the penalty contribution, and has 2201 parameters.
- Fitting the model for a single choice of λ's takes approximately 5 seconds.

Example 3: Selected Smooth Components



Example 3: Lasso Shrinkage of Linear Components



Lasso λ Selection – Another Approach

Lasso shrinkage is a special form of the TV penalty:

$$R_\tau(b) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b)$$

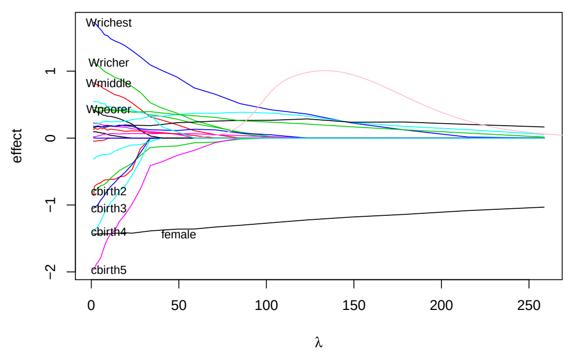
$$\begin{aligned} \hat\beta_{\tau,\lambda} &= \operatorname{argmin}\{R_\tau(b) + \lambda\|b\|_1\} \\ &\in \{b : 0 \in \partial R_\tau(b) + \lambda\partial\|b\|_1\}. \end{aligned}$$

At the true parameter, $\beta_0(\tau)$, we have the pivotal statistic,

$$\begin{aligned} \partial R_\tau(\beta_0(\tau)) &= \sum (\tau - I(F_{y_i}(y_i) \leqslant \tau))x_i \\ &\sim \sum (\tau - I(U_i \leqslant \tau))x_i \end{aligned}$$

Proposal: (Belloni and Chernozhukov (2009)) Choose λ as the $1-\alpha$ quantile of the simulated distribution of $\|\sum(\tau - I(U_i \leqslant \tau))x_i\|_\infty$ with iid $U_i \sim U[0,1]$.

Example 3: Lasso Shrinkage of Linear Components



Conclusions

- Nonparametric specifications of $Q(\tau|x)$ improve flexibility.
- Additive models keep effective dimension in check.
- Total variation roughness penalties are natural.
- Schwarz model selection criteria are useful for λ selection
- Hotelling tubes are useful for uniform confidence bands
- Lasso Shrinkage is useful for parametric components.

Endogoneity and All That

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011

Is there IV for QR?

- Amemiya (1982) and Powell (1983) consider analogues of 2SLS for median regression models
- Chen and Portnoy (1986) consider extensions to quantile regression
- Abadie, Angrist and Imbens (2002) consider models with binary endogonous treatment
- Chernozhukov and Hansen (2003) propose "inverse" quantile regression
- Chesher (2003) considers triangular models with continuous endogonous variables.

Chernozhukov and Hansen QRIV

Motivation: Yet another way to view two stage least squares.

Model: $y = X\beta + Z\alpha + u, \quad W \perp\!\!\!\perp u$

Estimator:

$$\hat{\alpha} = \text{argmin}_{\alpha} \|\hat{\gamma}(\alpha)\|^2_{A = W^\top M_X W}$$

$$\hat{\gamma}(\alpha) = \text{argmin}_{\gamma} \|y - X\beta - Z\alpha - W\gamma\|^2$$

Thm $\hat{\alpha} = (Z^\top P_{M_X W} Z)^{-1} Z^\top P_{M_X W} y$, the 2SLS estimator.

Heuristic: $\hat{\alpha}$ is chosen to make $\|\hat{\gamma}(\alpha)\|$ as small as possible to satisfy (approximately) the exclusion restriction/assumption.

Generalization: The quantile regression version simply replaces $\|\cdot\|^2$ in the definition of $\hat{\gamma}$ by the corresponding QR norm.

A Linear Location Shift Recursive Model

$$\begin{aligned} Y &= S\alpha_1 + x^\top \alpha_2 + \epsilon + \lambda\nu \qquad (1) \\ S &= z\beta_1 + x^\top \beta_2 + \nu \qquad (2) \end{aligned}$$

Suppose: $\epsilon \perp\!\!\!\perp \nu$ and $(\epsilon, \nu) \perp\!\!\!\perp (z, x)$. Substituting for ν from (2) into (1),

$$\begin{aligned} Q_Y(\tau_1 | S, x, z) &= S(\alpha_1 + \lambda) + x^\top(\alpha_2 - \lambda\beta_2) + z(-\lambda\beta_1) + F_\epsilon^{-1}(\tau_1) \\ Q_S(\tau_2 | z, x) &= z\beta_1 + x^\top \beta_2 + F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\begin{aligned} \pi_1(\tau_1, \tau_2) &= \nabla_{S_i} Q_{Y_i}|_{S_i = Q_{S_i}} + \frac{\nabla_{z_i} Q_{Y_i}|_{S_i = Q_{S_i}}}{\nabla_{z_i} Q_{S_i}} \\ &= (\alpha_1 + \lambda) + (-\lambda\beta_1)/\beta_1 \\ &= \alpha_1 \end{aligned}$$

A Linear Location-Scale Shift Model

$$\begin{aligned} Y &= S\alpha_1 + x^\top \alpha_2 + S(\epsilon + \lambda\nu) \\ S &= z\beta_1 + x^\top \beta_2 + \nu \\ \pi_1(\tau_1, \tau_2) &= \alpha_1 + F_\epsilon^{-1}(\tau_1) + \lambda F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\begin{aligned} Q_Y(\tau_1 | S, x, z) &= S\theta_1(\tau_1) + x^\top \theta_2 + S^2\theta_3 + Sz\theta_4 + Sx^\top \theta_5 \\ Q_S(\tau_2 | z, x) &= z\beta_1 + x^\top \beta_2 + F_\nu^{-1}(\tau_2) \end{aligned}$$

$$\hat{\pi}_1(\tau_1, \tau_2) = \sum_{i=1}^n w_i \left\{ \hat{\theta}_1(\tau_1) + 2\hat{Q}_{S_i}\hat{\theta}_3(\tau_1) + z_i\hat{\theta}_4(\tau_1) + x_i^\top \hat{\theta}_5(\tau_1) + \frac{\hat{Q}_{S_i}\hat{\theta}_4(\tau_1)}{\hat{\beta}_1(\tau_2)} \right\}$$

a weighted average derivative estimator with $\hat{Q}_{S_i} = \hat{Q}_S(\tau_2 | z_i, x_i)$.

The General Recursive Model

$$\begin{aligned} Y &= \varphi_1(S, x, \epsilon, \nu; \alpha) \\ S &= \varphi_2(z, x, \nu; \beta) \end{aligned}$$

Suppose: $\epsilon \perp\!\!\!\perp \nu$ and $(\epsilon, \nu) \perp\!\!\!\perp (z, x)$. Solving for ν and substituting we have the conditional quantile functions,

$$\begin{aligned} Q_Y(\tau_1 | S, x, z) &= h_1(S, x, z, \theta(\tau_1)) \\ Q_S(\tau_2 | z, x) &= h_2(z, x, \beta(\tau_2)) \end{aligned}$$

Extensions to more than two endogonous variables are "straightforward."

The (Chesher) Weighted Average Derivative Estimator

$$\begin{aligned}\hat{\theta}(\tau_1) &= \text{argmin}_\theta \sum_{i=1}^n \rho_{\tau_1}(Y_i - h_1(S, x, z, \theta(\tau_1))) \\ \hat{\beta}(\tau_2) &= \text{argmin}_\beta \sum_{i=1}^n \rho_{\tau_2}(S_i - h_2(z, x, \beta(\tau_2)))\end{aligned}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, giving structural estimators:

$$\hat{\pi}_1(\tau_1, \tau_2) = \sum_{i=1}^n w_i \Big\{ \nabla_S \hat{h}_{1i}|_{S_i = \hat{h}_{2i}} + \frac{\nabla_z \hat{h}_{1i}|_{S_i = \hat{h}_{2i}}}{\nabla_z \hat{h}_{2i}} \Big\},$$

$$\hat{\pi}_2(\tau_1, \tau_2) = \sum_{i=1}^n w_i \Big\{ \nabla_x \hat{h}_{1i}|_{S_i = \hat{h}_{2i}} - \frac{\nabla_z \hat{h}_{1i}|_{S_i = \hat{h}_{2i}}}{\nabla_z \hat{h}_{2i}} \nabla_x \hat{h}_{2i} \Big\},$$

2SLS as a Control Variate Estimator

$$\begin{aligned}Y &= S\alpha_1 + X_1\alpha_2 + u \equiv Z\alpha + u \\ S &= X\beta + V, \;\; \text{where} \;\; X = [X_1 \vdots X_2]\end{aligned}$$

Set $\hat{V} = S - \hat{S} \equiv M_X Y_1$, and consider the least squares estimator of the model,

$$Y = Z\alpha + \hat{V}\gamma + w$$

Claim: $\hat{\alpha}_{CV} \equiv (Z^\top M_{\hat{V}} Z)^{-1} Z^\top M_{\hat{V}} Y = (Z^\top P_X Z)^{-1} Z^\top P_X Y \equiv \hat{\alpha}_{2SLS}$.

Proof of Control Variate Equivalence

$$M_{\hat{V}} = M_{M_X S} = I - M_X S (S^\top M_X S)^{-1} S^\top M_X$$

$$\begin{aligned}S^\top M_{\hat{V}} &= S^\top - S^\top M_X = S^\top P_X \\ X_1^\top M_{\hat{V}} &= X_1^\top - X_1^\top M_X = X_1^\top = X_1^\top P_X\end{aligned}$$

Reward for information leading to a reference prior to Dhrymes (1970). Recent work on the control variate approach by Blundell, Powell, Smith, Newey and others.

Quantile Regression Control Variate Estimation I

Location scale shift model:

$$\begin{aligned}Y &= S(\alpha_1 + \epsilon + \lambda\nu) + x^\top \alpha_2 \\ S &= z\beta_1 + x^\top \beta_2 + \nu.\end{aligned}$$

Using $\hat{\nu}(\tau_2) = S - \hat{Q}_S(\tau_2|z, x)$ as a control variate,

$$\begin{aligned}Y &= w^\top \alpha(\tau_1, \tau_2) + \lambda S(\hat{Q}_S - Q_S) + S(\epsilon - F_\epsilon^{-1}(\tau_1)), \\ \text{where} \quad & w^\top = (S, x^\top, S\hat{\nu}(\tau_2)) \\ & \alpha(\tau_1, \tau_2) = (\alpha_1(\tau_1, \tau_2), \alpha_2, \lambda)^\top \\ & \alpha_1(\tau_1, \tau_2) = \alpha_1 + F_\epsilon^{-1}(\tau_1) + \lambda F_\nu^{-1}(\tau_2).\end{aligned}$$

$$\hat{\alpha}(\tau_1, \tau_2) = \text{argmin}_a \sum_{i=1}^n \rho_{\tau_1}(Y_i - w_i^\top a).$$

Quantile Regression Control Variate Estimation II

$$\begin{aligned}Y &= \varphi_1(S, x, \epsilon, \nu; \; \alpha) \\ S &= \varphi_2(z, x, \nu; \; \beta)\end{aligned}$$

Regarding $\nu(\tau_2) = \nu - F_\nu^{-1}(\tau_2)$ as a control variate, we have

$$\begin{aligned}Q_Y(\tau_1|S, x, \nu(\tau_2)) &= g_1(S, x, \nu(\tau_2), \alpha(\tau_1, \tau_2)) \\ Q_S(\tau_2|z, x) &= g_2(z, x, \beta(\tau_2))\end{aligned}$$

$$\hat{\nu}(\tau_2) = \varphi_2^{-1}(S, z, x, \hat{\beta}) - \varphi_2^{-1}(\hat{Q}_s, z, x, \hat{\beta})$$

$$\hat{\alpha}(\tau_1, \tau_2) = \text{argmin}_a \sum_{i=1}^n \rho_{\tau_1}(Y_i - g_1(S, x, \hat{\nu}(\tau_2), a)).$$

Asymptopia

Theorem: Under regularity conditions, the weighted average derivative and control variate estimators of the Chesher structural effect have an asymptotic linear (Bahadur) representation, and after efficient reweighting of both estimators, the control variate estimator has smaller covariance matrix than the weighted average derivative estimator.

Remark: The control variate estimator imposes more stringent restrictions on the estimation of the hybrid structural equation and should thus be expected to perform better when the specification is correct. The advantages of the control variate approach are magnified in situations of overidentification.

Asymptotics for WAD

Theorem

The $\hat{\pi}_n(\tau_1, \tau_2)$ has the asymptotic linear (Bahadur) representation,

$$\begin{aligned}\sqrt{n}(\hat{\pi}_n(\tau_1,\tau_2) - \pi(\tau_1,\tau_2)) = {} & W_1 \bar{J}_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_{i1} \dot{h}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) \\ & + W_2 \bar{J}_2^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_{i2} \dot{h}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) \\ & \Longrightarrow \mathcal{N}(0,\ \omega_{11} W_1 \bar{J}_1^{-1} J_1 \bar{J}_1^{-1} W_1^\top + \omega_{22} W_2 \bar{J}_2^{-1} J_2 \bar{J}_2^{-1} W_2^\top)\end{aligned}$$

$$J_j = \lim_{n\to\infty} \frac{1}{n} \sum \sigma_{ij}^2 \dot{h}_{ij} \dot{h}_{ij}^\top, \quad \bar{J}_j = \lim_{n\to\infty} \frac{1}{n} \sum \sigma_{ij} f_{ij}(\xi_{ij}) \dot{h}_{ij} \dot{h}_{ij}^\top,$$
$$W_1 = \nabla_\theta \pi(\tau_1,\tau_2), \quad W_2 = \nabla_\beta \pi(\tau_1,\tau_2),$$
$$\dot{h}_{i1} = \nabla_\theta h_{i1}, \quad \dot{h}_{i2} = \nabla_\beta h_{i2}, \quad \omega_{jj} = \tau_j(1-\tau_j).$$

Asymptotics for CV

Theorem

The $\hat{\alpha}_n(\tau_1, \tau_2)$ has the Bahadur representation,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}_n(\tau_1,\tau_2) - \alpha(\tau_1,\tau_2)) = {} & \bar{D}_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_{i1} \dot{g}_{i1} \psi_{\tau_1}(Y_{i1} - \xi_{i1}) \\ & + \bar{D}_1^{-1} \bar{D}_{12} \bar{D}_2^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_{i2} \dot{g}_{i2} \psi_{\tau_2}(Y_{i2} - \xi_{i2}) \\ & \Longrightarrow \mathcal{N}(0,\ \omega_{11} \bar{D}_1^{-1} D_1 \bar{D}_1^{-1} + \omega_{22} \bar{D}_1^{-1} \bar{D}_{12} \bar{D}_2^{-1} D_2 \bar{D}_2^{-1} \bar{D}_{12}^\top \bar{D}_1^{-1})\end{aligned}$$

$$D_j = \lim_{n\to\infty} n^{-1} \sum \sigma_{ij}^2 \dot{g}_{ij} \dot{g}_{ij}^\top, \quad \bar{D}_j = \lim_{n\to\infty} n^{-1} \sum \sigma_{ij} f_{ij}(\xi_{ij}) \dot{g}_{ij} \dot{g}_{ij}^\top,$$
$$\bar{D}_{12} = \lim_{n\to\infty} n^{-1} \sum \sigma_{i1} f_{i1} \eta_i \dot{g}_{i1} \dot{g}_{i2}^\top,$$
$$\dot{g}_{i1} = \nabla_\alpha g_{i1}, \quad \dot{g}_{i2} = \nabla_\beta g_{i2}, \quad \eta_i = (\partial g_{1i} / \partial \nu_{i2}(\tau_2))(\nabla_{\nu_{i2}} \varphi_{i2})^{-1}.$$

ARE of WAD and CV

- Efficient weights: $\sigma_{ij} = f_{ij}(\xi_{ij})$

$$\sqrt{n}(\hat{\pi}_n(\tau_1,\tau_2) - \pi(\tau_1,\tau_2)) \Rightarrow \mathcal{N}(0, \omega_{11} W_1 J_1^{-1} W_1^\top + \omega_{22} W_2 J_2^{-1} W_2^\top)$$
$$\sqrt{n}(\hat{\alpha}_n(\tau_1,\tau_2) - \alpha(\tau_1,\tau_2)) \Rightarrow \mathcal{N}(0, \omega_{11} D_1^{-1} + \omega_{22} D_1^{-1} D_{12} D_2^{-1} D_{12}^\top D_1^{-1}).$$

The mapping: $\tilde{\pi}_n = L\hat{\alpha}_n, \quad L\alpha = \pi.$

$$\begin{aligned} W_1 J_1^{-1} W_1^\top & \geqslant L D_1^{-1} L^\top \\ W_2 J_2^{-1} W_2^\top & \geqslant L D_1^{-1} D_{12} D_2^{-1} D_{12}^\top D_1^{-1} L^\top. \end{aligned}$$

Theorem

Under efficient reweighting of both estimators,

$$Avar(\sqrt{n}\tilde{\pi}_n) \leqslant Avar(\sqrt{n}\hat{\pi}_n).$$

Conclusions

- Triangular structural models facilitate causal analysis via recursive conditioning, directed acyclic graph representation.
- Recursive conditional quantile models yield interpretable heterogeneous structural effects.
- Control variate methods offer computationally and statistically efficient strategies for estimating heterogeneous structural effects.
- Weighted average derivative methods offer a less restrictive strategy for estimation that offers potential for model diagnostics and testing.

Censored Quantile Regression and Survival Models

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011

Quantile Regression for Duration (Survival) Models

A wide variety of survival analysis models, following Doksum and Gasko (1990), may be written as,

$$h(T_i) = x_i^\top \beta + u_i$$

where h is a monotone transformation, and

- T_i is an observed survival time,
- x_i is a vector of covariates,
- β is an unknown parameter vector
- $\{u_i\}$ are iid with df F.

The Cox Model

For the proportional hazard model with

$$\log \lambda(t|x) = \log \lambda_0(t) - x^\top \beta$$

the conditional survival function in terms of the integrated baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ as,

$$\log(-\log(S(t|x))) = \log \Lambda_0(t) - x^\top \beta$$

so, evaluating at $t = T_i$, we have the model,

$$\log \Lambda_0(T) = x^\top \beta + u$$

for u_i iid with df $F_0(u) = 1 - e^{-e^u}$.

The Bennett (Proportional-Odds) Model

For the proportional odds model, where the conditional odds of death $\Gamma(t|x) = F(t|x)/(1 - F(t|x))$ are written as,

$$\log \Gamma(t|x) = \log \Gamma_0(t) - x^\top \beta,$$

we have, similarly,

$$\log \Gamma_0(T) = x^\top \beta + u$$

for u iid logistic with $F_0(u) = (1 + e^{-u})^{-1}$.

Accelerated Failure Time Model

In the accelerated failure time model we have

$$\log(T_i) = x_i^\top \beta + u_i$$

so

$$\begin{aligned} P(T > t) &= P(e^u > te^{-x\beta}) \\ &= 1 - F_0(te^{-x\beta}) \end{aligned}$$

where $F_0(\cdot)$ denotes the df of e^u, and thus,

$$\lambda(t|x) = \lambda_0(te^{-x\beta})e^{-x\beta}$$

where $\lambda_0(\cdot)$ denotes the hazard function corresponding to F_0. In effect, the covariates act to rescale time in the baseline hazard.

Beyond the Transformation Model

The common feature of all these models is that after transformation of the observed survival times we have:

- a pure location-shift, iid-error regression model
- covariate effects shift the center of the distribution of $h(T)$, but
- covariates cannot affect scale, or shape of this distribution

An Application: Longevity of Mediterrean Fruit Flies

In the early 1990's there were a series of experiments designed to study the survival distribution of lower animals. One of the most influential of these was:

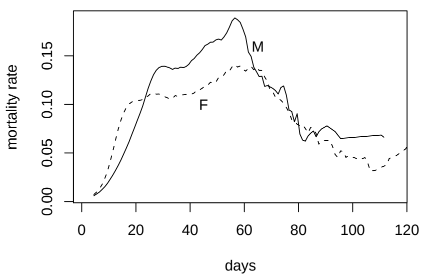
CAREY, J.R., LIEDO, P., OROZCO, D. AND VAUPEL, J.W. *(1992) Slowing of mortality rates at older ages in large Medfly cohorts, Science,* **258**, *457-61.*

- 1,203,646 medflies survival times recorded in days
- Sex was recorded on day of death
- Pupae were initially sorted into one of five size classes
- 167 aluminum mesh cages containing roughly 7200 flies
- Adults were given a diet of sugar and water *ad libitum*

Major Conclusions of the Medfly Experiment

- Mortality rates declined at the oldest observed ages. contradicting the traditional view that aging is an inevitable, monotone process of senescence.
- The right tail of the survival distribution was, at least by human standards, remarkably long.
- There was strong evidence for a crossover in gender specific mortality rates.

Lifetable Hazard Estimates by Gender



Smoothed mortality rates for males and females.

Medfly Survival Prospects

Lifespan (in days)	Percentage Surviving	Number Surviving
40	5	60,000
50	1	12,000
86	.01	120
146	.001	12

Initial Population of 1,203,646

Human Survival Prospects*

Lifespan (in years)	Percentage Surviving	Number Surviving
50	98	591,000
75	69	413,000
85	33	200,000
95	5	30,000
105	.08	526
115	.0001	1

* Estimated Thatcher (1999) Model

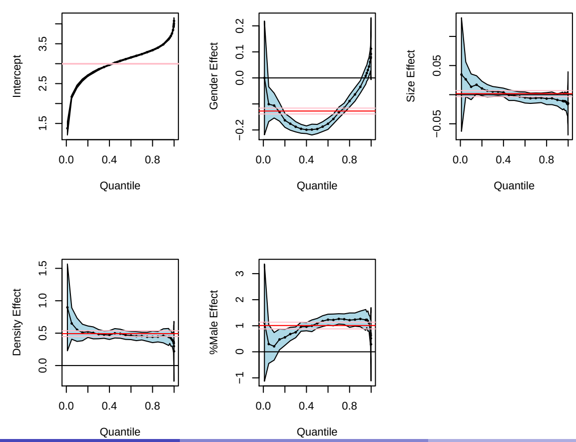
Quantile Regression Model (Geling and K (JASA,2001))

Criticism of the Carey et al paper revolved around whether declining hazard rates were a result of confounding factors of cage density and initial pupal size. Our basic QR model included the following covariates:

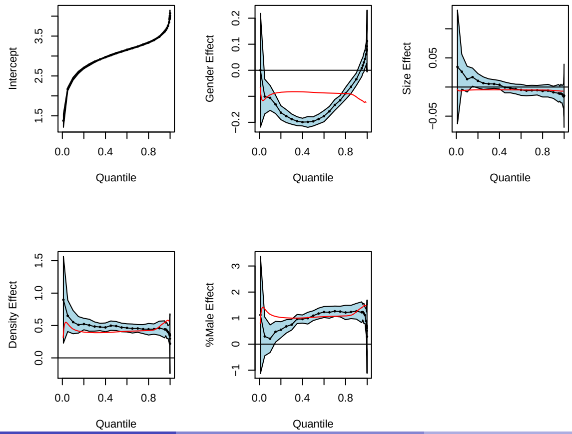
$$\begin{aligned} Q_{\log(T_i)}(\tau|x_i) &= \beta_0(\tau) + \beta_1(\tau)\mathrm{SEX} + \beta_2(\tau)\mathrm{SIZE} \\ &+ \beta_3(\tau)\mathrm{DENSITY} + \beta_4(\tau)\%\mathrm{MALE} \end{aligned}$$

- SEX Gender
- SIZE Pupal Size in mm
- DENSITY Initial Density of Cage
- %MALE Initial Proportion of Males

Base Model Results with AFT Fit



Base Model Results with Cox PH Fit



What About Censoring?

There are currently 3 approaches to handling censored survival data within the quantile regression framework:

- Powell (1986) Fixed Censoring
- Portnoy (2003) Random Censoring, Kaplan-Meier Analogue
- Peng/Huang (2008) Random Censoring, Nelson-Aalen Analogue

Available for R in the package `quantreg`.

Powell's Approach for Fixed Censoring

Rationale *Quantiles are equivariant to monotone transformation:*

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)) \text{ for } h \nearrow$$

Model $Y_i = T_i \wedge C_i \equiv \min\{T_i, C_i\}$

$$Q_{Y_i|x_i}(\tau|x_i) = x_i^\top \beta(\tau) \wedge C_i$$

Data *Censoring times are known for all observations*

$$\{Y_i, C_i, x_i : i = 1, \cdots, n\}$$

Estimator *Conditional quantile functions are nonlinear in parameters:*

$$\hat{\beta}(\tau) = \text{argmin} \sum \rho_\tau(Y_i - x_i^\top \beta \wedge C_i)$$

Portnoy's Approach for Random Censoring I

Rationale *Efron's (1967) interpretation of Kaplan-Meier as shifting mass of censored observations to the right:*

Algorithm *Until we "encounter" a censored observation KM quantiles can be computed by solving, starting at $\tau = 0$,*

$$\hat{\xi}(\tau) = \text{argmin}_\xi \sum_{i=1}^n \rho_\tau(Y_i - \xi)$$

Once we "encounter" a censored observation, i.e. when $\hat{\xi}(\tau_i) = y_i$ for some y_i with $\delta_i = 0$, we split y_i into two parts:

- $y_i^{(1)} = y_i$ *with weight* $w_i = (\tau - \tau_i)/(1 - \tau_i)$
- $y_i^{(2)} = y_\infty = \infty$ *with weight* $1 - w_i$.

Then denoting the index set of censored observations "encountered" up to τ by $K(\tau)$ we can solve

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - \xi) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - \xi) + (1 - w_i(\tau))\rho_\tau(y_\infty - \xi)].$$

Portnoy's Approach for Random Censoring II

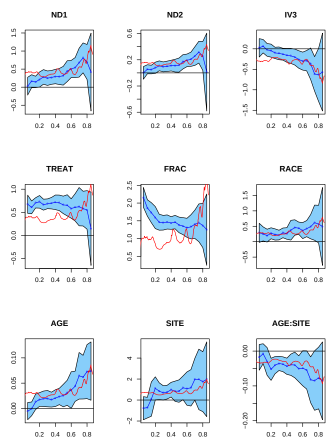
When we have covariates we can replace ξ by the inner product $x_i^\top \beta$ and solve:

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - x_i^\top \beta) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - x_i^\top \beta) + (1 - w_i(\tau))\rho_\tau(y_\infty - x_i^\top \beta)].$$

At each τ this is a simple, weighted linear quantile regression problem. The following R code fragment replicates an analysis in Portnoy (2003):

```
require(quantreg)
data(uis)
fit <- crq(Surv(log(TIME), CENSOR) ~ ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis, method = "Por")
Sfit <- summary(fit,1:19/20)
PHit <- coxph(Surv(TIME, CENSOR) ~ ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis)
plot(Sfit, CoxPHit = PHit)
```


Reanalysis of the Hosmer-Lemeshow Drug Relapse Data



Peng and Huang's Approach for Random Censoring I

Rationale *Extend the martingale representation of the Nelson-Aalen estimator of the cumulative hazard function to produce an "estimating equation" for conditional quantiles.*

Model *AFT form of the quantile regression model:*

$$\textit{Prob}(\log T_i \leqslant x_i^\top \beta(\tau)) = \tau$$

Data $\{(Y_i, \delta_i) : i = 1, \cdots, n\}$ $Y_i = T_i \wedge C_i, \ \delta_i = I(T_i < C_i)$

Martingale *We have* $EM_i(t) = 0$ *for* $t \geqslant 0$*, where:*

$$\begin{aligned} M_i(t) &= N_i(t) - \Lambda_i(t \wedge Y_i|x_i) \\ N_i(t) &= I(\{Y_i \leqslant t\}, \{\delta_i = 1\}) \\ \Lambda_i(t) &= -\log(1 - F_i(t|x_i)) \\ F_i(t) &= \textit{Prob}(T_i \leqslant t|x_i) \end{aligned}$$

Peng and Huang's Approach for Random Censoring II

The estimating equation becomes,

$$En^{-1/2} \sum x_i [N_i(\exp(x_i^\top \beta(\tau))) - \int_0^\tau I(Y_i \geqslant \exp(x_i^\top \beta(u))) dH(u) = 0.$$

where $H(u) = -\log(1 - u)$ for $u \in [0, 1)$, after rewriting:

$$\begin{aligned} \Lambda_i(\exp(x_i^\top \beta(\tau)) \wedge Y_i|x_i)) &= H(\tau) \wedge H(F_i(Y_i|x_i)) \\ &= \int_0^\tau I(Y_i \geqslant \exp(x_i^\top \beta(u))) dH(u), \end{aligned}$$

Approximating the integral on a grid, $0 = \tau_0 < \tau_1 < \cdots < \tau_J < 1$ yields a simple linear programming formulation to be solved at the gridpoints.

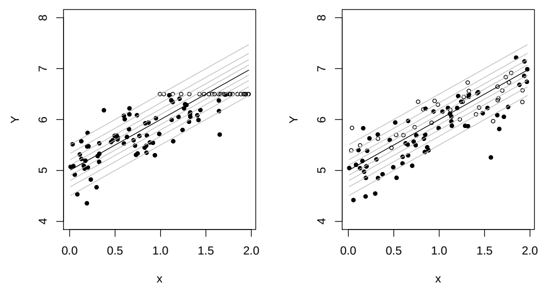
Alice in Asymptopia

It might be thought that the Powell estimator would be more efficient than the Portnoy and Peng-Huang estimators given that it imposes more stringent data requirements. Comparing asymptotic behavior and finite sample performance in the simplest one-sample setting indicates otherwise.

	median	Kaplan-Meier	Nelson-Aalen	Powell	Leurgans $\hat{G}$	Leurgans G
n= 50	1.602	1.972	2.040	2.037	2.234	2.945
n= 200	1.581	1.924	1.930	2.110	2.136	2.507
n= 500	1.666	2.016	2.023	2.187	2.215	2.742
n= 1000	1.556	1.813	1.816	2.001	2.018	2.569
n= ∞	1.571	1.839	1.839	2.017	2.017	2.463

Scaled MSE for Several Estimators of the Median: Mean squared error estimates are scaled by sample size to conform to asymptotic variance computations. Here, T_i is standard lognormal, and C_i is exponential with rate parameter .25, so the proportion of censored observations is roughly 30 percent. 1000 replications.

Simulation Settings I



Simulations I-A

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
n = 100	-0.0032	0.0638	0.0988	0.0025	0.0702	0.1063
n = 400	-0.0066	0.0406	0.0578	0.0036	0.0391	0.0588
n = 1000	-0.0022	0.0219	0.0321	0.0006	0.0228	0.0344
Peng-Huang						
n = 100	0.0005	0.0631	0.0986	0.0092	0.0727	0.1073
n = 400	-0.0007	0.0393	0.0575	0.0074	0.0389	0.0598
n = 1000	0.0014	0.0215	0.0324	0.0019	0.0226	0.0347
Powell						
n = 100	-0.0014	0.0694	0.1039	0.0068	0.0827	0.1252
n = 400	-0.0066	0.0429	0.0622	0.0098	0.0475	0.0734
n = 1000	-0.0008	0.0224	0.0339	0.0013	0.0264	0.0396
GMLE						
n = 100	0.0013	0.0528	0.0784	-0.0001	0.0517	0.0780
n = 400	-0.0039	0.0307	0.0442	0.0031	0.0264	0.0417
n = 1000	0.0003	0.0172	0.0248	-0.0001	0.0165	0.0242

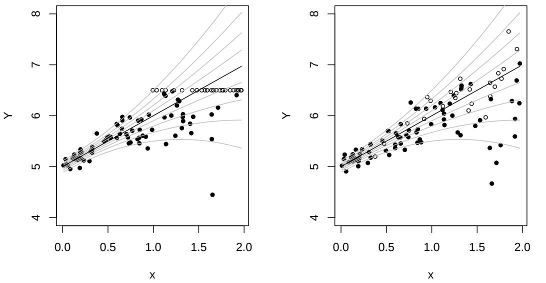
Comparison of Performance for the iid Error, Constant Censoring Configuration

Simulations I-B

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
n = 100	-0.0042	0.0646	0.0942	0.0024	0.0586	0.0874
n = 400	-0.0025	0.0373	0.0542	-0.0009	0.0322	0.0471
n = 1000	-0.0025	0.0208	0.0311	0.0006	0.0191	0.0283
Peng-Huang						
n = 100	0.0026	0.0639	0.0944	0.0045	0.0607	0.0888
n = 400	0.0056	0.0389	0.0547	-0.0002	0.0320	0.0476
n = 1000	0.0019	0.0212	0.0311	0.0009	0.0187	0.0283
Powell						
n = 100	-0.0025	0.0669	0.1017	0.0083	0.0656	0.1012
n = 400	0.0014	0.0398	0.0581	-0.0006	0.0364	0.0531
n = 1000	-0.0013	0.0210	0.0319	0.0016	0.0203	0.0304
GMLE						
n = 100	0.0007	0.0540	0.0781	0.0009	0.0470	0.0721
n = 400	0.0008	0.0285	0.0444	-0.0008	0.0253	0.0383
n = 1000	-0.0004	0.0169	0.0248	0.0002	0.0150	0.0224

Comparison of Performance for the iid Error, Variable Censoring Configuration

Simulation Settings II



Simulations II-A

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
n = 100	0.0084	0.0316	0.0396	-0.0251	0.0763	0.0964
n = 400	0.0076	0.0194	0.0243	-0.0247	0.0429	0.0533
n = 1000	0.0081	0.0121	0.0149	-0.0241	0.0309	0.0376
Portnoy Q						
n = 100	0.0018	0.0418	0.0527	0.0144	0.1576	0.2093
n = 400	-0.0010	0.0228	0.0290	0.0047	0.0708	0.0909
n = 1000	-0.0006	0.0122	0.0154	-0.0027	0.0463	0.0587
Peng-Huang L						
n = 100	0.0077	0.0313	0.0392	-0.0145	0.0749	0.0949
n = 400	0.0064	0.0193	0.0240	-0.0125	0.0392	0.0493
n = 1000	0.0077	0.0120	0.0147	-0.0181	0.0279	0.0342
Peng-Huang Q						
n = 100	0.0078	0.0425	0.0538	0.0483	0.1707	0.2328
n = 400	0.0035	0.0228	0.0291	0.0302	0.0775	0.1008
n = 1000	0.0015	0.0123	0.0155	0.0101	0.0483	0.0611
Powell						
n = 100	0.0021	0.0304	0.0385	-0.0034	0.0790	0.0993
n = 400	-0.0017	0.0191	0.0239	0.0028	0.0431	0.0544
n = 1000	-0.0001	0.0099	0.0125	0.0003	0.0257	0.0316
GMLE						
n = 100	0.1080	0.1082	0.1201	-0.2040	0.2042	0.2210
n = 400	0.1209	0.1209	0.1241	-0.2134	0.2134	0.2173
n = 1000	0.1118	0.1118	0.1130	-0.2075	0.2075	0.2091

Comparison of Performance for the Constant Censoring, Heteroscedastic Configuration

Simulations II-B

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
n = 100	0.0024	0.0278	0.0417	-0.0067	0.0690	0.1007
n = 400	0.0019	0.0145	0.0213	-0.0080	0.0333	0.0493
n = 1000	0.0016	0.0097	0.0139	-0.0062	0.0210	0.0312
Portnoy Q						
n = 100	0.0011	0.0352	0.0540	0.0094	0.1121	0.1902
n = 400	0.0002	0.0185	0.0270	-0.0012	0.0510	0.0774
n = 1000	-0.0005	0.0116	0.0169	-0.0011	0.0337	0.0511
Peng-Huang L						
n = 100	0.0018	0.0281	0.0417	0.0041	0.0694	0.1017
n = 400	0.0013	0.0142	0.0212	0.0035	0.0333	0.0490
n = 1000	0.0012	0.0096	0.0139	0.0002	0.0208	0.0310
Peng-Huang Q						
n = 100	0.0044	0.0364	0.0550	0.0322	0.1183	0.2105
n = 400	0.0026	0.0188	0.0275	0.0154	0.0504	0.0813
n = 1000	0.0007	0.0113	0.0169	0.0077	0.0333	0.0520
Powell						
n = 100	-0.0001	0.0288	0.0430	0.0055	0.0733	0.1105
n = 400	0.0000	0.0147	0.0226	0.0001	0.0379	0.0561
n = 1000	-0.0008	0.0095	0.0146	0.0013	0.0237	0.0350
GMLE						
n = 100	0.1078	0.1038	0.1272	-0.1576	0.1582	0.1862
n = 400	0.1123	0.1116	0.1168	-0.1581	0.1578	0.1647
n = 1000	0.1153	0.1138	0.1174	-0.1609	0.1601	0.1639

Comparison of Performance for the Variable Censoring, Heteroscedastic Configuration

Conclusions

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are quite similar.
- The martingale representation of the Peng-Huang estimator yields a more complete asymptotic theory than is currently available for the Portnoy estimator.
- The Powell estimator, although conceptually attractive, suffers from some serious computational difficulties, imposes strong data requirements, and has an inherent asymptotic efficiency disadvantage.
- Quantile regression provides a flexible complement to classical survival analysis methods, and is now well equipped to handle censoring.

Quantile Autoregression

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011

Based on joint work with Zhijie Xiao, Boston College.

Outline

Introduction

In classical regression and autoregression models

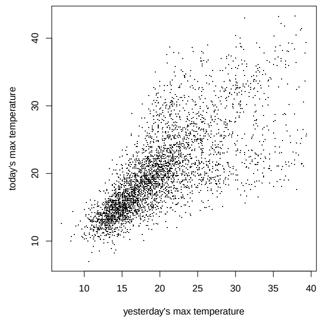
$$
\begin{aligned}
y_i &= h(x_i, \theta) + u_i, \\
y_t &= \alpha y_{t-1} + u_t
\end{aligned}
$$

conditioning covariates influence only the location of the conditional distribution of the response:

$$\text{Response} = \text{Signal} + \text{IID Noise}.$$

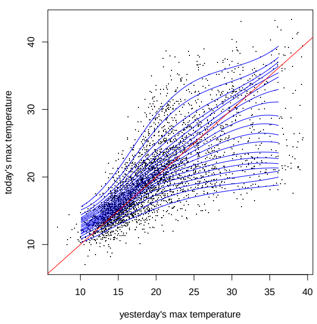
But why should noise always be so well-behaved?

A Motivating Example



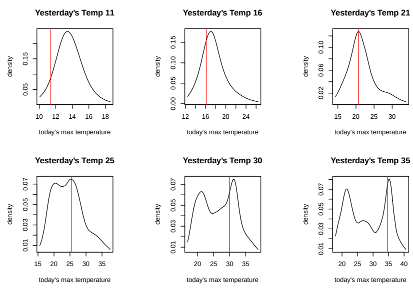
Daily Temperature in Melbourne: An AR(1) Scatterplot

Estimated Conditional Quantiles of Daily Temperature



Daily Temperature in Melbourne: A Nonlinear QAR(1) Model

Conditional Densities of Melbourne Daily Temperature



Location, scale and shape all change with y_{t-1}.
When today is hot, tomorrow's temperature is bimodal!

Linear AR(1) and QAR(1) Models

The classical linear AR(1) model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t,$$

with iid errors, $u_t : t = 1, \cdots, T$, implies

$$E(y_t | \mathcal{F}_{t-1}) = \alpha_0 + \alpha_1 y_{t-1}$$

and conditional quantile functions are all parallel:

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1 y_{t-1}$$

with $\alpha_0(\tau) = F_u^{-1}(\tau)$ just the quantile function of the u_t's.
But isn't this rather boring? What if we let α_1 depend on τ too?

A Random Coefficient Interpretation

If the conditional quantiles of the response satisfy:

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1}$$

then we can generate responses from the model by replacing τ by uniform random variables:

$$y_t = \alpha_0(u_t) + \alpha_1(u_t) y_{t-1} \quad u_t \sim \text{iid } U[0,1].$$

This is a very special form of random coefficient autoregressive (RCAR) model with comonotonic coefficients.

On Comonotonicity

Definition: Two random variables $X, Y : \Omega \to \mathbb{R}$ are comonotonic if there exists a third random variable $Z : \Omega \to \mathbb{R}$ and increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$.

- If X and Y are comonotonic they have rank correlation one.
- From our point of view the crucial property of comonotonic random variables is the behavior of quantile functions of their sums, X, Y comonotonic implies:

$$F_{X+Y}^{-1}(\tau) = F_X^{-1}(\tau) + F_Y^{-1}(\tau)$$

- X and Y are driven by the same random (uniform) variable.

The QAR(p) Model

Consider a p-th order QAR process,

$$Q_{y_t}(\tau | \mathcal{F}_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1} + \ldots + \alpha_p(\tau) y_{t-p}$$

Equivalently, we have random coefficient model,

$$\begin{aligned} y_t &= \alpha_0(u_t) + \alpha_1(u_t) y_{t-1} + \cdots + \alpha_p(u_t) y_{t-p} \\ &\equiv x_t^\top \alpha(u_t). \end{aligned}$$

Now, all $p+1$ random coefficients are comonotonic, functionally dependent on the same uniform random variable.

Vector QAR(1) representation of the QAR(p) Model

$$Y_t = \mu + A_t Y_{t-1} + V_t$$

where

$$\mu = \begin{bmatrix} \mu_0 \\ 0_{p-1} \end{bmatrix}, \; A_t = \begin{bmatrix} a_t & \alpha_p(u_t) \\ I_{p-1} & 0_{p-1} \end{bmatrix}, \; V_t = \begin{bmatrix} v_t \\ 0_{p-1} \end{bmatrix}$$

$$a_t = [\alpha_1(u_t), \ldots, \alpha_{p-1}(u_t)],$$

$$Y_t = [y_t, \cdots, y_{t-p+1}]^\top,$$

$$v_t = \alpha_0(u_t) - \mu_0.$$

It all looks rather complex and multivariate, but it is really still nicely univariate and very tractable.

Slouching Toward Asymptopia

We maintain the following regularity conditions:

A.1 $\{v_t\}$ are iid with mean 0 and variance $\sigma^2 < \infty$. The CDF of v_t, F, has a continuous density f with $f(v) > 0$ on $\mathcal{V} = \{v : 0 < F(v) < 1\}$.

A.2 Eigenvalues of $\Omega_A = E(A_t \otimes A_t)$ have moduli less than unity.

A.3 Denote the conditional CDF $\Pr[y_t < y | \mathcal{F}_{t-1}]$ as $F_{t-1}(y)$ and its derivative as $f_{t-1}(y)$, f_{t-1} is uniformly integrable on $\mathcal{V}$.

Stationarity

Theorem 1: Under assumptions A.1 and A.2, the QAR(p) process y_t is covariance stationary and satisfies a central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (y_t - \mu_y) \Rightarrow \mathcal{N}\left(0, \omega_y^2\right),$$

with

$$\begin{aligned} \mu_y &= \frac{\mu_0}{1 - \sum_{j=1}^{p} \mu_p}, \\ \mu_j &= E(\alpha_j(u_t)), \quad j = 0, ..., p, \\ \omega_y^2 &= \lim \frac{1}{n} E[\sum_{t=1}^{n} (y_t - \mu_y)]^2. \end{aligned}$$

Example: The QAR(1) Model

For the QAR(1) model,

$$Q_{y_t}(\tau | y_{t-1}) = \alpha_0(\tau) + \alpha_1(\tau) y_{t-1},$$

or with u_t iid $U[0, 1]$.

$$y_t = \alpha_0(u_t) + \alpha_1(u_t) y_{t-1},$$

if $\omega^2 = E(\alpha_1^2(u_t)) < 1$, then y_t is covariance stationary and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (y_t - \mu_y) \Rightarrow \mathcal{N}\left(0, \omega_y^2\right),$$

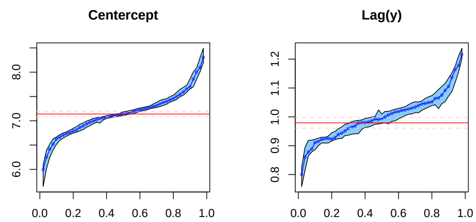
where $\mu_0 = E\alpha_0(u_t)$, $\mu_1 = E(\alpha_1(u_t))$, $\sigma^2 = V(\alpha_0(u_t))$, and

$$\mu_y = \frac{\mu_0}{(1 - \mu_1)}, \quad \omega_y^2 = \frac{(1 + \mu_1)\sigma^2}{(1 - \mu_1)(1 - \omega^2)},$$

Qualitative Behavior of QAR(p) Processes

- The model can exhibit unit-root-like tendencies, even temporarily explosive behavior, but episodes of mean reversion are sufficient to insure stationarity.
- Under certain conditions,the QAR(p) process is a semi-strong ARCH(p) process in the sense of Drost and Nijman (1993).
- The impulse response of y_{t+s} to a shock u_t is stochastic but converges (to zero) in mean square as $s \to \infty$.

Estimated QAR(1) v. AR(1) Models of U.S. Interest Rates



Data: Seasonally adjusted monthly: April, 1971 to June, 2002.
Do 3-month T-bills really have a unit root?

Estimation of the QAR model

Estimation of the QAR models involves solving,

$$\hat{\alpha}(\tau) = \text{argmin}_\alpha \sum_{t=1}^{n} \rho_\tau(y_t - x_t^\top \alpha),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, the $\surd$-function.
Fitted conditional quantile functions of y_t, are given by,

$$\hat{Q}_t(\tau | x_t) = x_t^\top \hat{\alpha}(\tau),$$

and conditional densities by the difference quotients,

$$\hat{f}_t(\tau | x_{t-1}) = \frac{2h}{\hat{Q}_t(\tau + h | x_{t-1}) - \hat{Q}_t(\tau - h | x_{t-1})},$$

The QAR Process

Theorem 2: Under our regularity conditions,

$$\sqrt{n}\Omega^{-1/2}(\hat{\alpha}(\tau) - \alpha(\tau)) \Rightarrow \mathcal{B}_{p+1}(\tau),$$

a $(p + 1)$-dimensional standard Brownian Bridge, with

$$\begin{aligned} \Omega &= \Omega_1^{-1} \Omega_0 \Omega_1^{-1}. \\ \Omega_0 &= E(x_t x_t^\top) = \lim n^{-1} \sum_{t=1}^{n} x_t x_t^\top, \\ \Omega_1 &= \lim n^{-1} \sum_{t=1}^{n} f_{t-1}(F_{t-1}^{-1}(\tau)) x_t x_t^\top. \end{aligned}$$

Inference for QAR models

For fixed $\tau = \tau_0$ we can test the hypothesis:

$$H_0: \quad R\alpha(\tau) = r$$

using the Wald statistic,

$$W_n(\tau) = \frac{n(R\hat{\alpha}(\tau) - r)^\top [R\hat{\Omega}_1^{-1}\hat{\Omega}_0\hat{\Omega}_1^{-1}R^\top]^{-1}(R\hat{\alpha}(\tau) - r)}{\tau(1-\tau)}$$

This approach can be extended to testing on general index sets $\tau \in \mathcal{T}$ with the corresponding Wald process.

Asymptotic Inference

Theorem: Under H_0, $W_n(\tau) \Rightarrow Q_m^2(\tau)$, where $Q_m(\tau)$ is a Bessel process of order $m = \text{rank}(R)$. For fixed τ, $Q_m^2(\tau) \sim \chi_m^2$.

- Kolmogorov-Smirov or Cramer-von-Mises statistics based on $W_n(\tau)$ can be used to implement the tests.
- For known R and r this leads to a very nice theory – estimated R and/or r testing raises new questions.
- The situation is quite analogous to goodness-of-fit testing with estimated parameters.

Example: Unit Root Testing

Consider the augmented Dickey-Fuller model

$$y_t = \delta_0 + \delta_1 y_{t-1} + \sum_{j=2}^{p} \delta_j \Delta y_{t-j} + u_t.$$

We would like to test this constant coefficients version of the model against the more general QAR(p) version:

$$Q_{y_t}(\tau|x_t) = \delta_0(\tau) + \delta_1(\tau)y_{t-1} + \sum_{j=2}^{p} \delta_j(\tau)\Delta y_{t-j}$$

The hypothesis: $H_0 : \delta_1(\tau) = \bar{\delta}_1 = 1$, for $\tau \in \mathcal{T} = [\tau_0, 1-\tau_0]$, is considered in Koenker and Xiao (JASA, 2004).

Example: Two Tests

- When $\bar{\delta}_1 < 1$ is known we have the candidate process,

  $$V_n(\tau) = \sqrt{n}(\hat{\delta}_1(\tau) - \bar{\delta}_1)/\hat{\omega}_{11}.$$

  where $\hat{\omega}_{11}^2$ is the appropriate element from $\hat{\Omega}_1^{-1}\hat{\Omega}_0\hat{\Omega}_1^{-1}$. Fluctuations in $V_n(\tau)$ can be evaluated with the Kolmogorov-Smirnov statistic,

  $$\sup_{\tau\in\mathcal{T}} V_n(\tau) \Rightarrow \sup_{\tau\in\mathcal{T}} B(\tau).$$

- When $\bar{\delta}_1$ is unknown we may replace it with an estimate, but this disrupts the convenient asymptotic behavior. Now,

  $$\hat{V}_n(\tau) = \sqrt{n}((\hat{\delta}_1(\tau) - \bar{\delta}_1) - (\hat{\delta}_1 - \bar{\delta}_1))/\hat{\omega}_{11}$$

Martingale Transformation of $\hat{V}_n(\tau)$

Khmaladze (1981) suggested a general approach to the transformation of parametric empirical processes like $\hat{V}_n(\tau)$:

$$\tilde{V}_n(\tau) = \hat{V}_n(\tau) - \int_0^\tau \left[\dot{g}_n(s)^\top C_n^{-1}(s)\int_s^1 \dot{g}_n(r)d\hat{V}_n(r)\right] ds$$

where $\dot{g}_n(s)$ and $C_n(s)$ are estimators of

$$\dot{g}(r) = (1, (\dot{f}/f)(F^{-1}(r)))^\top;\ C(s) = \int_s^1 \dot{g}(r)\dot{g}(r)^\top dr.$$

This is a generalization of the classical Doob-Meyer decomposition.

Restoration of the ADF property

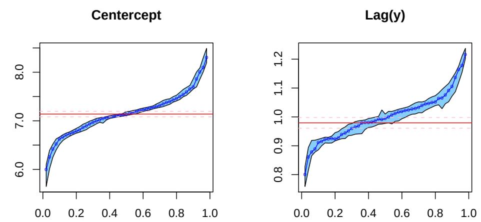
Theorem Under H_0, $\tilde{V}_n(\tau) \Rightarrow W(\tau)$ and therefore

$$\sup_{\tau\in\mathcal{T}} \|\tilde{V}_n(\tau)\| \Rightarrow \sup_{\tau\in\mathcal{T}} \|W(\tau)\|,$$

with $W(r)$ a standard Brownian motion.

- The martingale transformation of Khmaladze annihilates the contribution of the estimated parameters to the asymptotic behavior of the $\hat{V}_n(\tau)$ process, thereby restoring the asymptotically distribution free (ADF) character of the test.

Three Month T-Bills Again



A test of the "location-shift" hypothesis yields a test statistic of 2.76 which has a p-value of roughly 0.01, contradicting the conclusion of the conventional Dickey-Fuller test.

QAR Models for Longitudinal Data

- In estimating growth curves it is often valuable to condition not only on age, but also on prior growth and possibly on other covariates.
- Autoregressive models are natural, but complicated due to the irregular spacing of typical longitudinal measurements.
- Finnish Height Data: $\{Y_i(t_{i,j}) : j = 1, \dots, J_i, \ i = 1, \dots, n.\}$
- Partially Linear Model [Pere, Wei, Koenker, and He (2006)]:

$$Q_{Y_i(t_{i,j})}(\tau \mid t_{i,j}, Y_i(t_{i,j-1}), x_i) = g_\tau(t_{i,j}) + [\alpha(\tau) + \beta(\tau)(t_{i,j} - t_{i,j-1})]Y_i(t_{i,j-1}) + x_i^\top \gamma(\tau).$$

Parametric Components of the Conditional Growth Model

τ	Boys			Girls		
	$\hat\alpha(\tau)$	$\hat\beta(\tau)$	$\hat\gamma(\tau)$	$\hat\alpha(\tau)$	$\hat\beta(\tau)$	$\hat\gamma(\tau)$
0.03	0.845 (0.020)	0.147 (0.011)	0.024 (0.011)	0.809 (0.024)	0.135 (0.011)	0.042 (0.010)
0.1	0.787 (0.020)	0.159 (0.007)	0.036 (0.007)	0.757 (0.022)	0.153 (0.007)	0.054 (0.009)
0.25	0.725 (0.019)	0.170 (0.006)	0.051 (0.009)	0.685 (0.021)	0.163 (0.006)	0.061 (0.008)
0.5	0.635 (0.025)	0.173 (0.009)	0.060 (0.013)	0.612 (0.027)	0.175 (0.008)	0.070 (0.009)
0.75	0.483 (0.029)	0.187 (0.009)	0.063 (0.017)	0.457 (0.027)	0.183 (0.012)	0.094 (0.015)
0.9	0.422 (0.024)	0.213 (0.016)	0.070 (0.017)	0.411 (0.030)	0.201 (0.015)	0.100 (0.018)
0.97	0.383 (0.024)	0.214 (0.016)	0.077 (0.018)	0.400 (0.038)	0.232 (0.024)	0.086 (0.027)

Estimates of the QAR(1) parameters, $\alpha(\tau)$ and $\beta(\tau)$ and the mid-parental height effect, $\gamma(\tau)$, for Finnish children ages 0 to 2 years.

Forecasting with QAR Models

Given an estimated QAR model,

$$\hat Q_{y_t}(\tau | \mathcal{F}_{t-1}) = x_t^\top \hat\alpha(\tau)$$

based on data: $y_t : \ t = 1, 2, \cdots, T$, we can forecast

$$\hat y_{T+s} = \tilde x_{T+s}^\top \hat\alpha(U_s), \ s = 1, \cdots, S,$$

where $\tilde x_{T+s} = [1, \tilde y_{T+s-1}, \cdots, \tilde y_{T+s-p}]^\top$, $U_s \sim U[0,1]$, and

$$\tilde y_t = \begin{cases} y_t & \text{if} \quad t \leqslant T, \\ \hat y_t & \text{if} \quad t > T. \end{cases}$$

Conditional density forecasts can be made based on an ensemble of such forecast paths.

Linear QAR Models May Pose Statistical Health Risks

- Lines with distinct slopes eventually intersect. [Euclid: P5]
- Quantile functions, $Q_Y(\tau|x)$ should be monotone in τ for all x, intersections imply point masses – or even worse.
- What is to be done?
  - Constrained QAR: Quantiles can be estimated simultaneously subject to linear inequality restrictions.
  - Nonlinear QAR: Abandon linearity in the lagged y_t's, as in the Melbourne temperature example, both parametric and nonparametric options are available.

Nonlinear QAR Models via Copulas

An interesting class of stationary, Markovian models can be expressed in terms of their copula functions:

$$G(y_t, y_{t-1}, \cdots, y_{y-p}) = C(F(y_t), F(y_{t-1}), \cdots, F(y_{y-p}))$$

where G is the joint df and F the common marginal df.

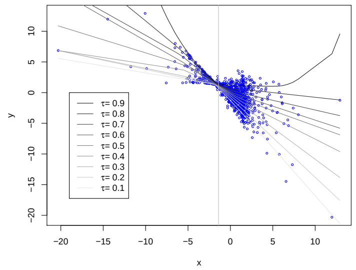
- Differentiating, $C(u, v)$, with respect to u, gives the conditional df,

$$H(y_t | y_{t-1}) = \frac{\partial}{\partial u} C(u, v)|_{(u=F(y_t), v=F(y_{t-1}))}$$

- Inverting we have the conditional quantile functions,

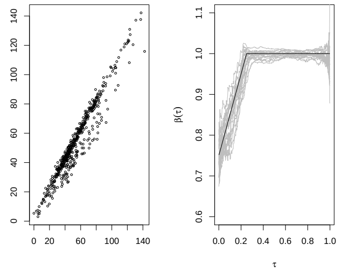
$$Q_{y_t}(\tau | y_{t-1}) = h(y_{t-1}, \theta(\tau))$$

Example 1 (Fan and Fan)



Model: $Q_{y_t}(\tau|y_{t-1}) = -(1.7 - 1.8\tau)y_{t-1} + \Phi^{-1}(\tau)$.

Example 2 (Near Unit Root)



Model: $Q_{y_t}(\tau|y_{t-1}) = 2 + \min\{\frac{3}{4} + \tau, 1\}y_{t-1} + 3\Phi^{-1}(\tau)$.

Conclusions

- QAR models are an attempt to expand the scope of classical linear time-series models permitting lagged covariates to influence scale and shape as well as location of conditional densities.
- Efficient estimation via familiar linear programming methods.
- Random coefficient interpretation nests many conventional models including ARCH.
- Wald-type inference is feasible for a large class of hypotheses; rank based inference is also an attractive option.
- Forecasting conditional densities is potentially valuable.
- Many new and challenging open problems. . . .

Quantile Regression for Longitudinal Data

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011

Classical Linear Fixed/Random Effects Model

Consider the model,

$$y_{ij} = x_{ij}^\top \beta + \alpha_i + u_{ij} \quad j = 1, \dots m_i, \quad i = 1, \dots, n,$$

or

$$y = X\beta + Z\alpha + u.$$

The matrix Z represents an incidence matrix that identifies the n distinct individuals in the sample. If u and α are independent Gaussian vectors with $u \sim \mathcal{N}(0, R)$ and $\alpha \sim \mathcal{N}(0, Q)$. Observing that $v = Z\alpha + u$ has covariance matrix $Evv^\top = R + ZQZ^\top$, we can immediately deduce that the minimum variance unbiased estimator of β is,

$$\hat{\beta} = (X^\top (R + ZQZ^\top)^{-1} X)^{-1} X^\top (R + ZQZ^\top)^{-1} y.$$

A Penalty Interpretation of $\hat{\beta}$

Proposition. $\hat{\beta}$ solves $\min_{(\alpha,\beta)} \|y - X\beta - Z\alpha\|^2_{R^{-1}} + \|\alpha\|^2_{Q^{-1}}$, where $\|x\|^2_A = x^\top A x$.

Proof.
Differentiating we obtain the normal equations,

$$X^\top R^{-1} X \hat{\beta} + X^\top R^{-1} Z \hat{\alpha} = X^\top R^{-1} y$$

$$Z^\top R^{-1} X \hat{\beta} + (Z^\top R^{-1} Z + Q^{-1}) \hat{\alpha} = Z^\top R^{-1} y$$

Solving, we have $\hat{\beta} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y$ where

$$\Omega^{-1} = R^{-1} - R^{-1} Z (Z^\top R^{-1} Z + Q^{-1})^{-1} Z^\top R^{-1}.$$

But $\Omega = R + ZQZ^\top$, see e.g. Rao(1973, p 33.). $\square$

This result has a long history: Henderson(1950), Goldberger(1962), Lindley and Smith (1972), etc.

Quantile Regression with Fixed Effects

Suppose that the conditional quantile functions of the response of the jth observation on the ith individual y_{ij} takes the form:

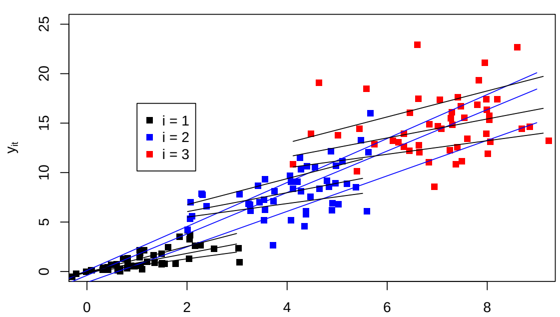
$$Q_{y_{ij}}(\tau | x_{ij}) = \alpha_i + x_{ij}^\top \beta(\tau) \quad j = 1, \dots m_i, \quad i = 1, \dots, n.$$

In this formulation the α's have a pure location shift effect on the conditional quantiles of the response. The effects of the covariates, x_{ij} are permitted to depend upon the quantile, τ, of interest, but the α's do not. To estimate the model for several quantiles simultaneously, we propose solving,

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n} \sum_{i=1}^{m_i} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^\top \beta(\tau_k))$$

Note that the usual between/within transformations are not permitted.

Penalized Quantile Regression with Fixed Effects



Time invariant, individual specific intercepts are quantile independent; slopes are quantile dependent.

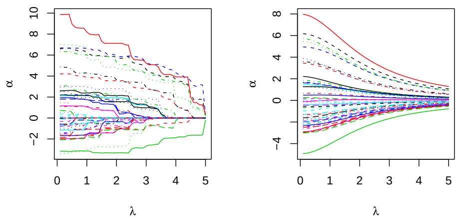
Penalized Quantile Regression with Fixed Effects

When n is large relative to the m_i's shrinkage may be advantageous in controlling the variability introduced by the large number of estimated α parameters. We will consider estimators solving the penalized version,

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n} \sum_{i=1}^{m_i} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^\top \beta(\tau_k)) + \lambda \sum_{i=1}^{n} |\alpha_i|.$$

For $\lambda \to 0$ we obtain the fixed effects estimator described above, while as $\lambda \to \infty$ the $\hat{\alpha}_i \to 0$ for all $i = 1, 2, \dots, n$ and we obtain an estimate of the model purged of the fixed effects. In moderately large samples this requires sparse linear algebra. Example R code is available from my webpages.

Shrinkage of the Fixed Effects



Shrinkage of the fixed effect parameter estimates, $\hat{\alpha}_i$. The left panel illustrates an example of the ℓ_1 shrinkage effect. The right panel illustrates an example of the ℓ_2 shrinkage effect.

Dynamic Panel Models and IV Estimation

Galvao (2010) considers dynamic panel models of the form:

$$Q_{y_{it}}(\tau | y_{i,t-1}, x_{it}) = \alpha_i + \gamma(\tau) y_{i,t-1} + x_{it}^\top \beta(\tau) \; t = 1, \ldots T_i, \; i = 1, \ldots, n.$$

In "short" panels estimation suffers from the same bias problems as seen in least squares estimators Nickel (1981) Hsiao and Anderson (1981); using the IV estimation approach of Chernozhukov and Hansen (2004) this bias can be reduced.

Correlated Random Effects

Abrevaya and Dahl (JBES, 2008) adapt the Chamberlain (1982) correlated random effects model and estimate a model of birthweight like that of Koenker and Hallock (2001).
The R package `rqpd` implements both this method and the penalized fixed effect approach. Available from R-Forge with the command:

```
install.packages("rqpd", repos="http://R-Forge.R-project.org")
```

This is a challenging, but very important, problem and hopefully there will be new and better approaches in the near future.

Risk, Choquet Portfolios and Quantile Regression

Roger Koenker

CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011

Joint work with Gib Bassett (UIC) and Gregory Kordas (Athens)

Outline

- Is there a useful role for pessimism in decision theory?
- A pessimistic theory of risk
- How to be pessimistic?

St. Petersburg Paradox

What would you be willing to pay to play the game:

$$G = \{\text{pay: } p, \quad \text{win: } 2^n \text{ with probability } 2^{-n}, \quad n = 1, 2, \ldots\}$$

Daniel Bernoulli ($\sim$ 1728) observed that even though the expected payoff was infinite, the gambler who maximized logarithmic utility would pay only a finite value to play. For example, given initial wealth 100,000 Roubles, our gambler would be willing to pay only 17 Roubles and 55 kopecks. If initial wealth were only 1000 Roubles, then the value of the game is only about 11 Roubles.

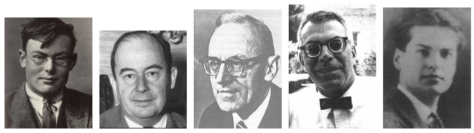
Expected Utility

To decide between two real valued gambles

$$X \sim F \quad \text{and} \quad Y \sim G$$

we choose X over Y if

$$Eu(X) = \int u(x)dF(x) \geqslant \int u(y)dG(y) = Eu(Y)$$



On Axiomatics

Suppose we have acts $P, Q, R, \ldots$ in a space $\mathcal{P}$, which admits enough convex structure to allow us to consider mixtures,

$$\alpha P + (1-\alpha)Q \in \mathcal{P} \quad \alpha \in (0,1)$$

Think of P, Q, R as probability measures on some underlying outcome/event space, $\mathcal{X}$.
Or better, view P, Q, R as acts mapping a space $\mathcal{S}$ of soon-to-be-revealed "states of nature" to the space of probability measures on the outcome space, $\mathcal{X}$.

The Expected Utility Theorem

Theorem(von-Neumann-Morgenstern) Suppose we have a preference relation $\{\succeq, \succ, \sim\}$ on $\mathcal{P}$ satisfying the axioms:

(A.1) (weak order) For all $P, Q, R \in \mathcal{P}$, $P \succeq Q$ or $Q \succeq P$, and $P \succeq Q$ and $Q \succeq R \Rightarrow P \succeq R$,

(A.2) (independence) For all $P, Q, R \in \mathcal{P}$ and $\alpha \in (0,1)$, then $P \succ Q \Rightarrow \alpha P + (1-\alpha)R \succ \alpha Q + (1-\alpha)R$,

(A.3) (continuity) For all $P, Q, R \in \mathcal{P}$, if $P \succ Q$ and $Q \succ R$, then there exist α and $\beta \in (0,1)$, such that, $\alpha P + (1-\alpha)R \succ \beta Q(1-\beta)R$.

Then there exists a linear function u on $\mathcal{P}$ such that for all $P, Q \in \mathcal{P}$, $P \succ Q$ if and only if $u(P) > u(Q)$.

Weakening the Independence Axiom

The independence axiom seems quite innocuous, but it is extremely powerful. We will consider a weaker form of independence due to Schmeidler (1989).

(A.2') (comonotonic independence) For all pairwise comonotonic $P, Q, R \in \mathcal{P}$ and $\alpha \in (0,1)$ $P \succ Q \Rightarrow \alpha P + (1-\alpha)R \succ \alpha Q + (1-\alpha)R$,

Definition Two acts P and Q in $\mathcal{P}$ are comonotonic, or similarly ordered, if for no s and t in $\mathcal{S}$,

$$P(\{t\}) \succ P(\{s\}) \quad \text{and} \quad Q(\{s\}) \succ Q(\{t\}).$$

"If P is better in state t than state s, then Q is also better in t than s."

On Comonotonicity

Definition The two functions $X, Y : \Omega \to \mathbb{R}$ are comonotonic if there exists a third function $Z : \Omega \to \mathfrak{R}$ and increasing functions f and g such that $X = f(Z)$ and $Y = g(Z)$.

From our point of view the crucial property of comonotonic random variables is the behavior of quantile functions of their sums. For comonotonic random variables X, Y, we have

$$F_{X+Y}^{-1}(u) = F_X^{-1}(u) + F_Y^{-1}(u)$$

By comonotonicity we have a $U \sim U[0,1]$ such that $Z = g(U) = F_X^{-1}(U) + F_Y^{-1}(U)$ where g is left continuous and increasing, so by monotone invariance, $F_{g(U)}^{-1} = g \circ F_U^{-1} = F_X^{-1} + F_Y^{-1}$. Comonotonic random variables are maximally dependent *a la* Fréchet

Choquet Expected Utility

Among the many proposals offered to extend expected utility theory the most attractive (to us) replaces

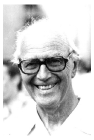
$$E_F u(X) = \int_0^1 u(F^{-1}(t))dt \geqslant \int_0^1 u(G^{-1}(t))dt = E_G u(Y)$$

with

$$E_{\nu,F} u(X) = \int_0^1 u(F^{-1}(t))d\nu(t) \geqslant \int_0^1 u(G^{-1}(t))d\nu(t) = E_{\nu,G} u(Y)$$

The measure ν permits distortion of the probability assessments after ordering the outcomes. This rank dependent form of expected utility has been pioneered by Quiggin (1981), Yaari (1987), Schmeidler (1989), Wakker (1989) and Dennenberg (1990).

Pessimism



By relaxing the independence axiom we obtain a larger class of preferences representable as Choquet capacities and introducing pessimism. The simplest form of Choquet expected utility is based on the "distortion"

$$\nu_\alpha(t) = \min\{t/\alpha, 1\}$$

so

$$E_{\nu_\alpha,F} u(X) = \alpha^{-1} \int_0^\alpha u(F^{-1}(t))dt$$

This exaggerates the probability of the proportion α of least favorable events, and totally discounts the probability of the $1-\alpha$ most favorable events.

Expect the worst – and you won't be disappointed.

A Smoother example

A simple, yet intriguing, one-parameter family of pessimistic Choquet distortions is the measure:

$$\nu_\theta(t) = 1 - (1-t)^\theta \qquad \theta \geqslant 1$$

Note that, changing variables, $t \to F_X(u)$, we have,

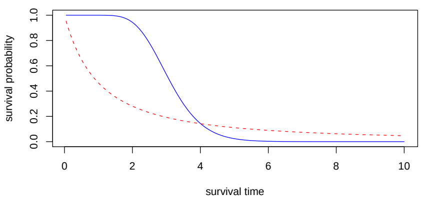
$$E_{\nu_\theta} X = \int_0^1 F_X^{-1}(t)d\nu(t) = \int_{-\infty}^{\infty} u d(1-(1-F_X(u))^\theta)$$

The pessimist imagines that he gets not a single draw from X but θ draws, and from these he always gets the worst. The parameter θ is a natural "measure of pessimism," and need not be an integer.

Savage on Pessimism

I have, at least once heard it objected against the personalistic view of probability that, according to that view, two people might be of different opinions, according as one is pessimistic and the other optimistic. I am not sure what position I would take in abstract discussion of whether that alleged property of personalistic views would be objectionable, but I think it is clear from the formal definition of qualitative probability that the particular personalistic view sponsored here does not leave room for optimism and pessimism, however these traits may be interpreted, to play any role in the person's judgement of probabilities. (Savage(1954), p. 68)

Pessimistic Medical Decision Making?



Survival Functions for a hypothetical medical treatment: The Lehmann quantile treatment effect (QTE) is the horizontal distance between the survival curves. In this example consideration of the mean treatment effect would slightly favor the (dotted) treatment curve, but the pessimistic patient might favor the (solid) placebo curve. Only the luckiest 15% actually do better under the treatment.

Risk as Pessimism?

In expected utility theory risk is entirely an attribute of the utility function:

Risk Neutrality	$\Rightarrow$	$u(x) \sim$ affine
Risk Aversion	$\Rightarrow$	$u(x) \sim$ concave
Risk Attraction	$\Rightarrow$	$u(x) \sim$ convex

Locally, the risk premium, i.e. the amount one is willing to pay to accept a zero mean risk, X, is

$$\pi(w, X) = \tfrac{1}{2}A(w)V(X)$$

where $A(w) = -u''(w)/u'(w)$ is the Arrow-Pratt coefficient of absolute risk aversion and $V(X)$ is the variance of X. Why is variance a reasonable measure of risk?

A Little Risk Aversion is a Dangerous Thing

Would you accept the gamble:

$$G_1 \qquad 50-50 \left\langle \begin{array}{l} \text{win } \$110 \\ \text{lose } \$100 \end{array} \right.$$

Suppose you say "no", then what about the gamble:

$$G_2 \qquad 50-50 \left\langle \begin{array}{l} \text{win } \$700,000 \\ \text{lose } \$1,000 \end{array} \right.$$

If you say "no" to G_1 for any initial wealth up to \$300,000, then you must also say "no" to G_2.
Moral: A little local risk aversion over small gambles implies implausibly large risk aversion over large gambles. Reference: Rabin (2000)

Are Swiss Bicycle Messengers Risk Averse?

When Veloblitz and Flash bicycle messengers from Zurich were confronted with the bet:

$$50-50 \left\langle \begin{array}{l} \text{win 8 CHF} \\ \text{lose 5 CHF} \end{array} \right.$$

More than half (54%) rejected the bet.
Reference: Fehr and Götte (2002)

Coherent Risk

Definition (Artzner, Delbaen, Eber and Heath (1999)) For real valued random variables $X \in \mathcal{X}$ on $(\Omega, \mathcal{A})$ a mapping $\rho : \mathcal{X} \to \mathcal{R}$ is called a coherent risk measure if,

1. Monotone: $X, Y \in \mathcal{X}$, with $X \leqslant Y \Rightarrow \rho(X) \geqslant \rho(Y)$.
2. Subadditive: $X, Y, X+Y \in \mathcal{X}, \Rightarrow \rho(X+Y) \leqslant \rho(X) + \rho(Y)$.
3. Linearly Homogeneous: For all $\lambda \geqslant 0$ and $X \in \mathcal{X}$, $\rho(\lambda X) = \lambda\rho(X)$.
4. Translation Invariant: For all $\lambda \in \mathcal{R}$ and $X \in \mathcal{X}$, $\rho(\lambda + X) = \rho(X) - \lambda$.

Many conventional measures of risks including those based on standard deviation are ruled out by these requirements. So are quantile based measures like "value at risk."

Choquet α-Risk

The leading example of a coherent risk measure is

$$\rho_{\nu_\alpha}(X) = -\alpha^{-1}\int_0^\alpha F^{-1}(t)dt$$

Variants of this risk measure have been introduced under several names

- Expected shortfall (Acerbi and Tasche (2002))
- Conditional VaR (Rockafellar and Uryasev (2000))
- Tail conditional expectation (Artzner, et al (1999)).

Note that $\rho_{\nu_\alpha}(X) = -E_{\nu_\alpha,F}(X)$, so Choquet α-risk is just negative Choquet expected utility with the distortion function ν_α.

Pessimistic Risk Measures

Definition A risk measure ρ will be called pessimistic if, for some probability measure φ on [0, 1]

$$\rho(X) = \int_0^1 \rho_{\nu_\alpha}(X) d\varphi(\alpha)$$

By Fubini

$$\begin{aligned}\rho(X) &= -\int_0^1 \alpha^{-1} \int_0^\alpha F^{-1}(t) dt d\varphi(\alpha) \\ &= -\int_0^1 F^{-1}(t) \int_t^1 \alpha^{-1} d\varphi(\alpha) dt \\ &\equiv -\int_0^1 F^{-1}(t) d\nu(t)\end{aligned}$$

Approximating General Pessimistic Risk Measures

We can approximate any pessimistic risk measure by taking

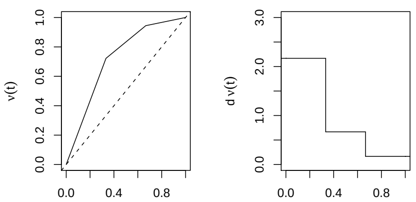
$$d\varphi(t) = \sum \varphi_i \delta_{\tau_i}(t)$$

where δ_τ denotes (Dirac) point mass 1 at τ. Then

$$\rho(X) = -\varphi_0 F^{-1}(0) - \int_0^1 F^{-1}(t)\gamma(t)dt$$

where $\gamma(t) = \sum \varphi_i \tau_i^{-1} I(t < \tau_i)$ and $\varphi_i > 0$, with $\sum \varphi_i = 1$.

An Example



$$d\varphi(t) = \tfrac{1}{2}\delta_{1/3}(t) + \tfrac{1}{3}\delta_{2/3}(t) + \tfrac{1}{6}\delta_1(t)$$

A Theorem

Theorem (Kusuoka (2001)) A regular risk measure is *coherent* in the sense of Artzner *et. al.* if and only if it is *pessimistic*.

- Pessimistic Choquet risk measures correspond to *concave* ν, i.e., *monotone decreasing* $d\nu$.
- Probability assessments are distorted to accentuate the probability of the least favorable events.
- The crucial coherence requirement is subadditivity, or submodularity, or 2-alternatingness in the terminology of Choquet capacities.

An Example

Samuelson (1963) describes asking a colleague at lunch whether he would be willing to make a

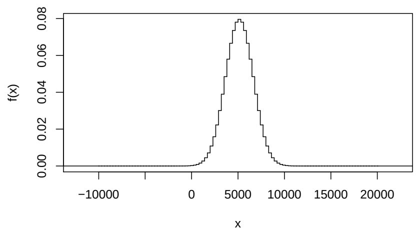
$$50 - 50 \text{ bet} \left\langle \begin{array}{l} \text{win } 200 \\ \text{lose } 100 \end{array} \right.$$

The colleague (later revealed to be E. Cary Brown) responded

"no, but I *would* be willing to make 100 such bets."

This response has been interpreted not only as reflecting a basic confusion about how to maximize expected utility but also as a fundamental misunderstanding of the law of large numbers.

Payoff Density of 100 Samuelson trials



Odds of losing money on the 100 trial bet is 1 chance in 2300.

Was Brown really irrational?

Suppose, for the sake of simplicity that

$$d\varphi(t) = \tfrac{1}{2}\delta_{1/2}(t) + \tfrac{1}{2}\delta_1(t)$$

so for one Samuelson coin flip we have the unfavorable evaluation,

$$E_{\nu,F}(X) = \tfrac{1}{2}(-100) + \tfrac{1}{2}(50) = -25$$

but for $S = \sum_{i=1}^{100} X_i \sim \mathcal{B}in(.5, 100)$ we have the favorable evaluation,

$$\begin{aligned} E_{\nu,F}(S) &= \tfrac{1}{2}2\int_0^{1/2} F_S^{-1}(t)dt + \tfrac{1}{2}(5000) \\ &= 1704.11 + 2500 \\ &= 4204.11 \end{aligned}$$

How to be Pessimistic

Theorem Let X be a real-valued random variable with $EX = \mu < \infty$, and $\rho_\alpha(u) = u(\alpha - I(u < 0))$. Then

$$\min_{\xi \in \mathbb{R}} E\rho_\alpha(X - \xi) = \alpha\mu + \rho_{\nu_\alpha}(X)$$

So α risk can be estimated by the sample analogue

$$\hat{\rho}_{\nu_\alpha}(x) = (n\alpha)^{-1} \min_{\xi} \sum \rho_\alpha(x_i - \xi) - \hat{\mu}_n$$

I knew it! Eventually everything looks like quantile regression to this guy!

Pessimistic Portfolios

Now let $X = (X_1, \ldots, X_p)$ denote a vector of potential portfolio asset returns and $Y = X^\top \pi$, the returns on the portfolio with weights π. Consider

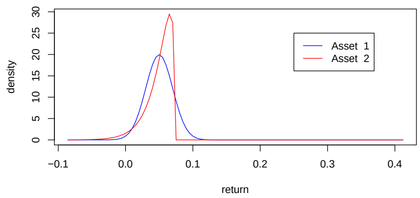
$$\min_{\pi} \rho_{\nu_\alpha}(Y) - \lambda\mu(Y)$$

Minimize α-risk subject to a constraint on mean return.
This problem can be formulated as a linear quantile regression problem

$$\min_{(\beta,\xi) \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\alpha(x_{i1} - \sum_{j=2}^{p}(x_{i1} - x_{ij})\beta_j - \xi) \quad \text{s.t.} \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

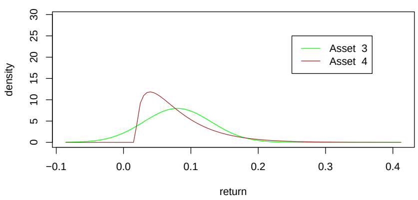
where $\pi(\beta) = (1 - \sum_{j=2}^{p} \beta_j, \beta^\top)^\top$.

An Example



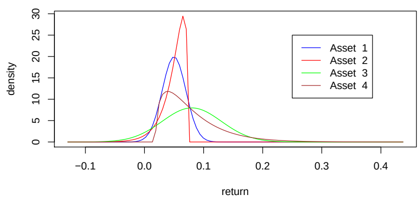
Two asset return densities with identical mean and variance.

An Example



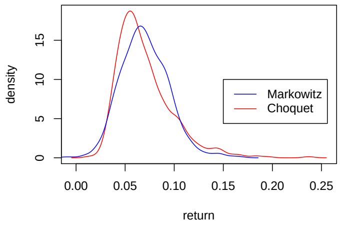
Two more asset return densities with identical mean and variance.

An Example



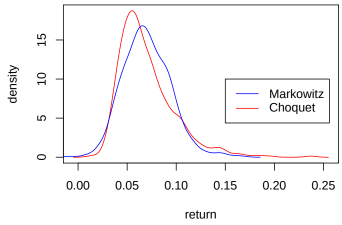
Two pairs of asset return densities with identical mean and variance.

Optimal Choquet and Markowitz Portfolio Returns



Markowitz portfolio minimizes the standard deviation of returns subject to mean return $\mu = .07$. The Choquet portfolio minimizes Choquet risk (for $\alpha = .10$) subject to earning the same mean return. The Choquet portfolio has better performance in both tails than mean-variance Markowitz portfolio.

Optimal Choquet and Markowitz Portfolio Returns



Now, the Markowitz portfolio minimizes the standard deviation of returns subject to mean return $\mu = .07$. The Choquet portfolio maximizes expected return subject to achieving the same Choquet risk (for $\alpha = .10$) as the Markowitz portfolio. Choquet portfolio has expected return $\mu = .08$ a full percentage point higher than the Markowitz portfolio.

A Unified Theory of Pessimism

Any pessimistic risk measure may be approximated by

$$\rho_\nu(X) = \sum_{k=1}^{m} \varphi_k \rho_{\nu_{\alpha_k}}(X)$$

where $\varphi_k > 0$ for $k = 1, 2, \ldots, m$ and $\sum \varphi_k = 1$.
Portfolio weights can be estimated for these risk measures by solving linear programs that are weighted sums of quantile regression problems:

$$\min_{(\beta,\xi)\in\mathcal{R}^p} \sum_{k=1}^{m}\sum_{i=1}^{n} \nu_k \rho_{\alpha_k}(x_{i1} - \sum_{j=2}^{p}(x_{i1} - x_{ij})\beta_j - \xi_k) \quad \text{s.t.} \quad \bar{x}^\top \pi(\beta) = \mu_0,$$

Software in R is available on from my web pages.

Conclusions

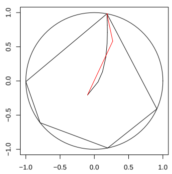
- Expected Utility is unsatisfactory both as a positive, i.e., descriptive, theory of behavior and as a normative guide to behavior.
- Choquet (non-additive, rank dependent) expected utility provides a simple, tractable alternative.
- Mean-variance Portfolio allocation is also unsatisfactory since it relies on unpalatable assumptions of Gaussian returns, or quadratic utility.
- Choquet portfolio optimization can be formulated as a quantile regression problem thus providing an attractive practical alternative to the dominant mean-variance approach of Markowitz (1952).

Quantile Regression Computation: From the Inside and the Outside

Roger Koenker

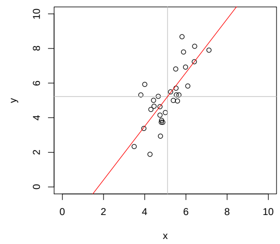
CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011



The Origin of Regression – Regression Through the Origin

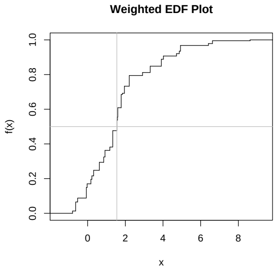
Find the line with mean residual zero that minimizes the sum of absolute residuals.



Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta|$ s.t. $\bar{y} = \alpha + \bar{x}\beta$.

Boscovich/Laplace *Methode de Situation*

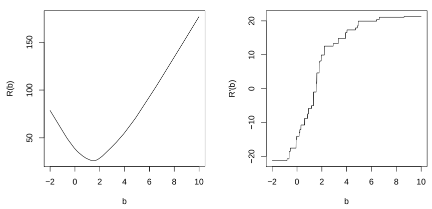
Algorithm: Order the n candidate slopes: $b_i = (y_i - \bar{y})/(x_i - \bar{x})$ denoting them by $b_{(i)}$ with associated weights $w_{(i)}$ where $w_i = |x_i - \bar{x}|$. Find the weighted median of these slopes.



Methode de Situation via Optimization

$$R(b) = \sum |\tilde{y}_i - \tilde{x}_i b| = \sum |\tilde{y}_i/\tilde{x}_i - b| \cdot |\tilde{x}_i|.$$

$$R'(b) = -\sum \mathrm{sgn}(\tilde{y}_i/\tilde{x}_i - b) \cdot |\tilde{x}_i|.$$



Quantile Regression through the Origin in R

This can be easily generalized to compute quantile regression estimates:

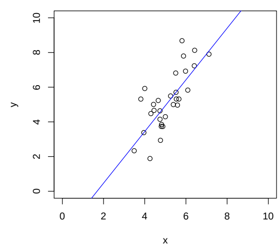
```
wquantile <- function(x, y, tau = 0.5) {
        o <- order(y/x)
        b <- (y/x)[o]
        w <- abs(x[o])
        k <- sum(cumsum(w) < ((tau - 0.5) * sum(x) + 0.5 * sum(w)))
        list(coef = b[k + 1], k = ord[k+1])
}
```

Warning: When $\bar{x} = 0$ then τ is irrelevant. Why?

Edgeworth's (1888) Plural Median

What if we want to estimate both α and β by median regression?

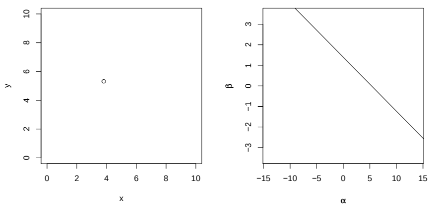
Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta|$



Edgeworth's (1888) Dual Plot: Anticipating Simplex

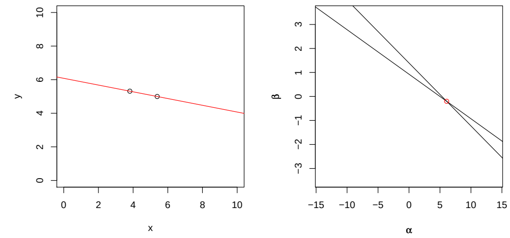
Points in sample space map to lines in parameter space.

$$(x_i, y_i) \mapsto \{(\alpha, \beta) : \alpha = y_i - x_i\beta\}$$



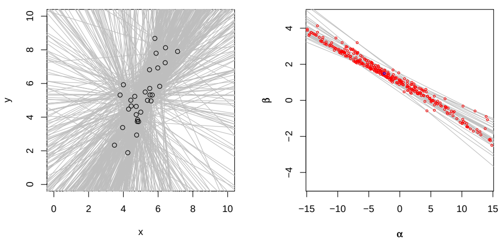
Edgeworth's (1888) Dual Plot: Anticipating Simplex

Lines through pairs of points in sample space map to points in parameter space.



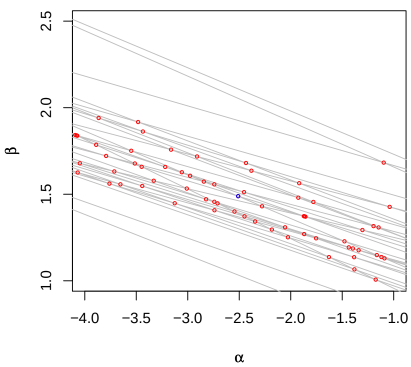
Edgeworth's (1888) Dual Plot: Anticipating Simplex

All pairs of observations produce $\binom{n}{2}$ points in dual plot.



Edgeworth's (1888) Dual Plot: Anticipating Simplex

Follow path of steepest descent through points in the dual plot.



Barrodale-Roberts Implementation of Edgeworth

```
rqx<- function(x, y, tau = 0.5, max.it = 50) { # Barrodale and Roberts -- lite
        p <- ncol(x); n <- nrow(x)
        h <- sample(1:n, size = p) #Phase I -- find a random (!) initial basis
        it <- 0
        repeat {
                it <- it + 1
                Xhinv <- solve(x[h,  ])
                bh <- Xhinv %*% y[h]
                rh <- y - x %*% bh
        #find direction of steepest descent along one of the edges
                g <-  - t(Xhinv) %*% t(x[ - h,  ]) %*% c(tau - (rh[ - h] < 0))
                g <- c(g + (1 - tau),  - g + tau)
                ming <- min(g)
                if(ming >= 0 || it > max.it) break
                h.out <- seq(along = g)[g == ming]
                sigma <- ifelse(h.out <= p, 1, -1)
                if(sigma < 0) h.out <- h.out - p
                d <- sigma * Xhinv[, h.out]
        #find step length by one-dimensional wquantile minimization
                xh <- x %*% d
                step <- wquantile(xh, rh, tau)
                h.in <- step$k
                h <- c(h[ - h.out], h.in)
        }
        if(it > max.it) warning("non-optimal solution: max.it exceeded")
        return(bh)
}
```


Linear Programming Duality

Primal: $\min_x\{c^\top x | Ax - b \in T,\ x \in S\}$
Dual: $\max_y\{b^\top y | c - A^\top y \in S^*,\ y \in T^*\}$

The sets S and T are closed convex cones, with dual cones S^* and T^*. A cone K^* is dual to K if:

$$K^* = \{y \in \mathbb{R}^n | x^\top y \geqslant 0 \text{ if } x \in K\}$$

Note that for any feasible point (x, y)

$$b^\top y \leqslant y^\top Ax \leqslant c^\top x$$

while optimality implies that

$$b^\top y = c^\top x.$$

Quantile Regression Primal and Dual

Splitting the QR "residual" into positive and negative parts, yields the primal linear program,

$$\min_{(b,u,v)} \{\tau 1^\top u + (1-\tau)1^\top v \mid Xb + u - v - y \in \{0\}, \quad (b,u,v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}\}.$$

with dual program:

$$\max_d \{y^\top d \mid X^\top d \in \{0\}, \quad \tau 1 - d \in \mathbb{R}_+^n, \quad (1-\tau)1 + d \in \mathbb{R}_+^n\},$$

$$\max_d \{y^\top d \mid X^\top d = 0, \; d \in [\tau - 1, \tau]^n\},$$

$$\max_a \{y^\top a \mid X^\top a = (1-\tau)X^\top 1, \quad a \in [0,1]^n\}$$

Quantile Regression Dual

The dual problem for quantile regression may be formulated as:

$$\max_a \{y^\top a \mid X^\top a = (1-\tau)X^\top 1, \; a \in [0,1]^n\}$$

What do these $\hat{a}_i(\tau)$'s mean statistically?
They are regression rank scores (Gutenbrunner and Jurečková (1992)):

$$\hat{a}_i(\tau) \in \begin{cases} \{1\} & \text{if} \quad y_i > x_i^\top \hat{\beta}(\tau) \\ (0,1) & \text{if} \quad y_i = x_i^\top \hat{\beta}(\tau) \\ \{0\} & \text{if} \quad y_i < x_i^\top \hat{\beta}(\tau) \end{cases}$$

The integral $\int \hat{a}_i(\tau) d\tau$ is something like the rank of the ith observation. It answers the question: On what quantile does the ith observation lie?

Linear Programming: The Inside Story

The Simplex Method (Edgeworth/Dantzig/Kantorovich) moves from vertex to vertex on the outside of the constraint set until it finds an optimum.
Interior point methods (Frisch/Karmarker/et al) take Newton type steps toward the optimal vertex from inside the constraint set.
A toy problem: Given a polygon inscribed in a circle, find the point on the polygon that maximizes the sum of its coordinates:

$$\max\{e^\top u \mid A^\top x = u, \; e^\top x = 1, \; x \geqslant 0\}$$

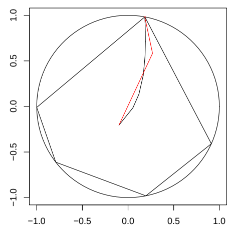
were e is vector of ones, and A has rows representing the n vertices.
Eliminating u, setting $c = Ae$, we can reformulate the problem as:

$$\max\{c^\top x \mid e^\top x = 1, \quad x \geqslant 0\},$$

Toy Story: From the Inside

Simplex goes around the outside of the polygon; interior point methods tunnel from the inside, solving a sequence of problems of the form:

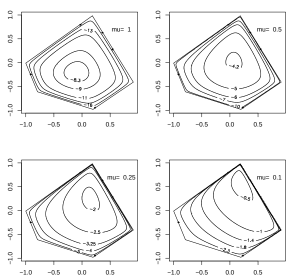
$$\max\{c^\top x + \mu \sum_{i=1}^n \log x_i \mid e^\top x = 1\}$$



Toy Story: From the Inside

By letting $\mu \to 0$ we get a sequence of smooth problems whose solutions approach the solution of the LP:

$$\max\{c^\top x + \mu \sum_{i=1}^n \log x_i \mid e^\top x = 1\}$$



Implementation: Meketon's Affine Scaling Algorithm

```
meketon <- function (x, y, eps = 1e-04, beta = 0.97) {
   f <- lm.fit(x,y)
   n <- length(y)
   w <- rep(0, n)
   d <- rep(1, n)
   its <- 0
   while(sum(abs(f$resid)) - crossprod(y, w) > eps) {
       its <- its + 1
       s <- f$resid * d
       alpha <- max(pmax(s/(1 - w), -s/(1 + w)))
       w <- w + (beta/alpha) * s
       d <- pmin(1 - w, 1 + w)^2
       f <- lm.wfit(x,y,d)
       }
   list(coef = f$coef, iterations = its)
   }
```

Mehrotra Primal-Dual Predictor-Corrector Algorithm

The algorithms implemented in `quantreg` for R are based on Mehrotra's Predictor-Corrector approach. Although somewhat more complicated than Meketon this has several advantages:

- Better numerical stability and efficiency due to better central path following,
- Easily generalized to incorporate linear inequality constraints.
- Easily generalized to exploit sparsity of the design matrix.

These features are all incorporated into various versions of the algorithm in `quantreg`, and coded in Fortran.

Back to Basics

Which is easier to compute: the median or the mean?

```
> x <- rnorm(100000000) # n = 10^8
> system.time(mean(x))
   user  system elapsed
 10.277   0.035  10.320
> system.time(kuantile(x,.5))
   user  system elapsed
  5.372   3.342   8.756
```

`kuantile` is a `quantreg` implementation of the Floyd-Rivest (1975) algorithm. For the median it requires $1.5n + O((n \log n)^{1/2})$ comparisons.

Portnoy and Koenker (1997) propose a similar strategy for "preprocessing" quantile regression problems to improve efficiency for large problems.

Globbing for Median Regression

Rather than solving $\min \sum |y_i - x_i b|$ consider:

1. Preliminary estimation using random $m = n^{2/3}$ subset,
2. Construct confidence band $x_i^\top \hat{\beta} \pm \kappa \|\hat{V}^{1/2} x_i\|$.
3. Find $J_L = \{i | y_i \text{ below band }\}$, and $J_H = \{i | y_i \text{ above band }\}$,
4. Glob observations together to form pseudo observations:

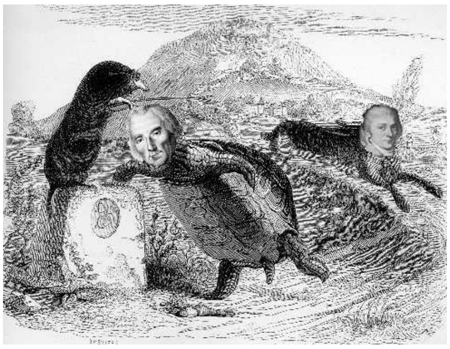
$$(x_L, y_L) = (\sum_{i \in J_L} x_i, -\infty), \quad (x_H, y_H) = (\sum_{i \in J_H} x_i, +\infty)$$

5. Solve the problem (with m+2 observations)

$$\min \sum |y_i - x_i b| + |y_L - x_L b| + |y_H - x_H b|$$

6. Verify that globbed observations have the correct predicted signs.

The Laplacian Tortoise and the Gaussian Hare



Retouched 18th century woodblock photo-print