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 four 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 probably unrealistic to expect polished answers to these questions by noon on Saturday, 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 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 many operating systems including MS Windows. 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) Venables and Ripley (2002). On the CRAN website there are also, under the heading "contributed", introductions to R in French, German, Spanish and Italian.

For purposes of this course a minimal knowledge of R will suffice. R will be available on the Citrix machine and should appear as an icon on the desktop of the laboratory machines. Clicking the icon should produce a window in which R will be running. To quit R, you must 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 "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 helpwindow should pop up and you can type more general key words or phrases to search for functionality. It is frequently helpful to save R commands into a file and execute a

These exercises were 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, and to the NSF for continuing research support under grant SES-0240781.

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.

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

```
> d <- read.table(file = "engel.data", header = TRUE)
> engel <- data.frame(d)
> attach(engel)
```

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:

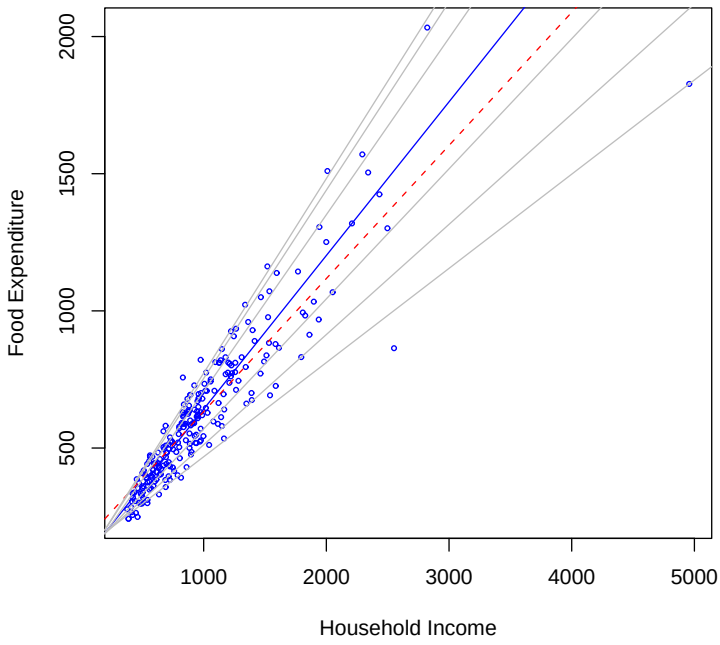
```
> plot(x, y)
```

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

```
> library(quantreg)
```

```
[1] "quantreg library loaded"
```

```
> plot(x, y, cex = 0.25, type = "n", xlab = "Household Income",
+     ylab = "Food Expenditure")
> points(x, y, cex = 0.5, col = "blue")
> abline(rq(y ~ x, tau = 0.5), col = "blue")
> abline(lm(y ~ x), lty = 2, col = "red")
> taus <- c(0.05, 0.1, 0.25, 0.75, 0.9, 0.95)
> for (i in 1:length(taus)) {
+     abline(rq(y ~ x, tau = taus[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 very slow in large applications. In `rq()` there are also lots of 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`.

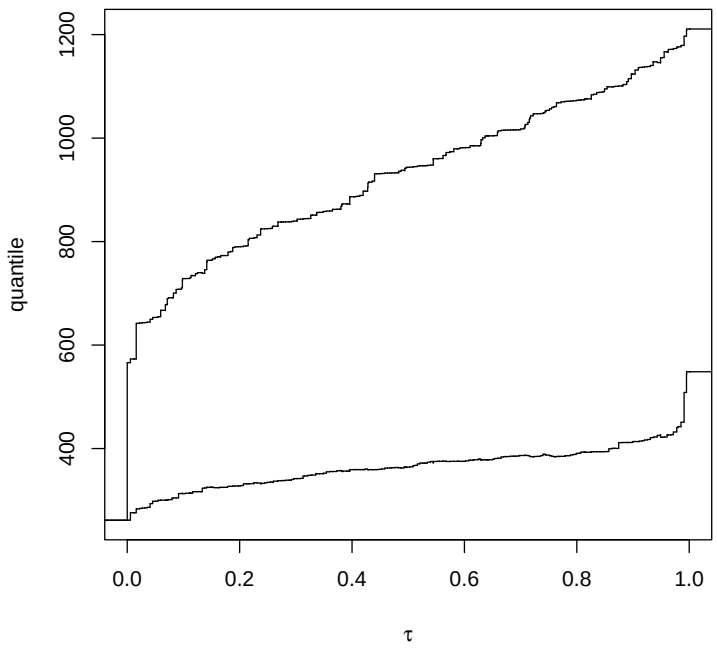
4. If we wanted to see all the distinct quantile regression solutions for this example we could specify a tau outside the range [0,1], e.g.

```
> 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 same 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:

```
> x.poor <- quantile(x, 0.1)
> x.rich <- quantile(x, 0.9)
> 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, ])
> plot(c(ps, ps), c(qs.poor, qs.rich), type = "n", xlab = expression(tau),
+     ylab = "quantile")
> library(stepfun)
> 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)
```



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 = 0.25)
> fit.25

Call:
rq(formula = y ~ x, tau = 0.25)

Coefficients:
            coefficients   lower bd    upper bd
(Intercept)   95.4835396 73.7860765 120.0984745
x              0.4741032  0.4203298   0.4943288

Degrees of freedom: 235 total; 233 residual

> fit.75 <- rq(y ~ x, tau = 0.75)
> fit.75

Call:
rq(formula = y ~ x, tau = 0.75)

Coefficients:
            coefficients   lower bd    upper bd
(Intercept)   62.3965855 32.7448768 107.3136214
x              0.6440141  0.5801552   0.6904127

Degrees of freedom: 235 total; 233 residual
```

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 Variance Table

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

   Df Resid Df F value    Pr(>F)
1   1      468  30.891 4.591e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

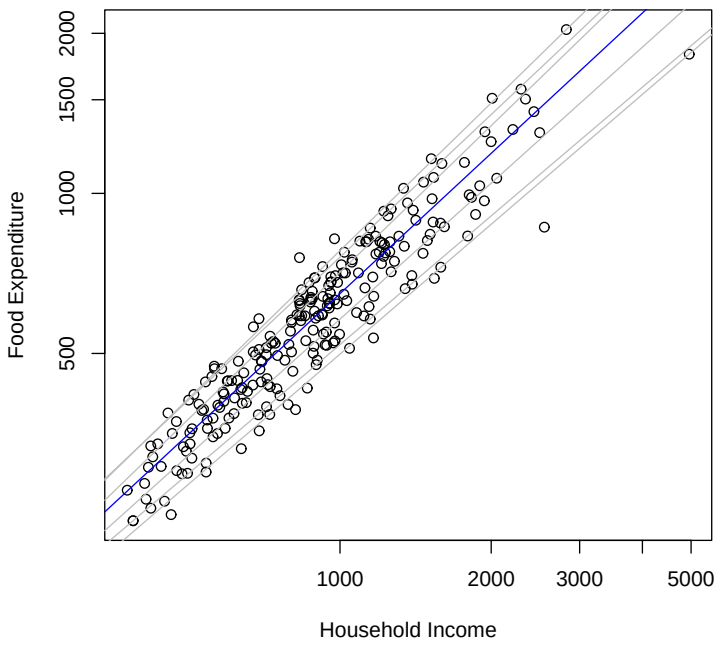
This is an example of a general class of tests proposed in Koenker and Bassett (1982) It is instructive to look at the code for the command `anova.rq` to see how this test is carried out. The Wald approach is used and the asymptotic covariance matrix is estimated using the approach of Hendricks and Koenker (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, no standard errors or confidence intervals are provided, only the point estimates are returned. (This effect is also possible to achieve with the default `method="br"` setting by adding the flag `ci=FALSE`. Details of the algorithms are provide in Koenker and d'Orey (1987), Koenker and d'Orey (1993),and Portnoy and Koenker (1997). In either case given the fit one can get standard inference results by calling summary, e.g.

```
> fit <- rq(y ~ x, tau = 0.27, method = "fn")
> summary(fit)

Call: rq(formula = y ~ x, tau = 0.27, method = "fn")

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 94.18652 20.15611    4.67285  0.00001
x            0.48321  0.02685   17.99854  0.00000
```

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.

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(0.05, 0.1, 0.25, 0.75, 0.9, 0.95)
> abline(rq(log10(y) ~ log10(x), tau = 0.5), 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 is carried out by the following function:

```
> "lprq" <- function(x, y, h, m = 50, tau = 0.5) {
+     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, ci = FALSE)
+         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:

```
> library(splines)
> fit <- rq(y ~ bs(x, df = 5), tau = 0.33)
```

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 = 0.33)
```

Compare some fitting using the spline approach with that obtained with the local polynomial kernel approach.

4. Another 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. Software to implement these methods is available from my website. 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 library, Koenker and Ng (2003).

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`. 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 `library(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: 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`. The R function `qrisk` computes optimal portfolio weights based on a matrix of observed, or simulated, asset returns using a specified form of pessimistic Choquet preferences.

PROBLEM 5: 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 Doksum and Aaberge. I would love to see a further applications along the lines outlined here.

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.