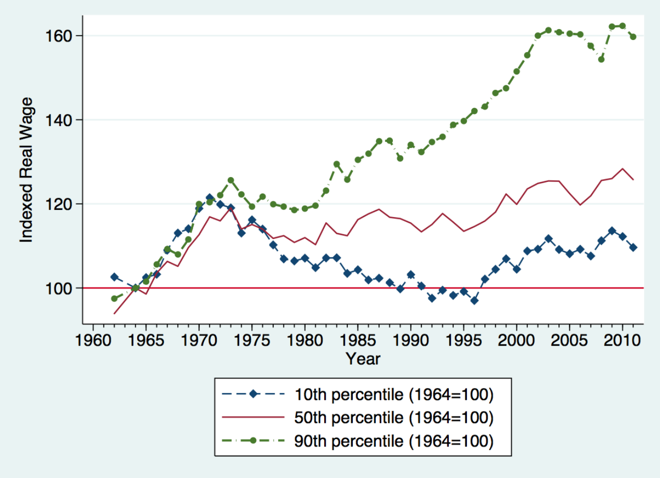
Week 15: Quantile Regression

JJ Chen

April 24, 2015

Motivation


Indexed Real Wage
160
140
120
100
1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010
Year
10th percentile (1964=100)
50th percentile (1964=100)
90th percentile (1964=100)

Motivation

- So far we've studied a lot about the CEF

$$\mathrm{Y}_i = \mathrm{E}(\mathrm{Y}_i \mid \mathrm{X}_i) + \varepsilon_i$$

- For continuous outcomes (or discrete variables with many values), we might want to know what happens to the whole distribution
  - Job training programs: Earnings
  - Obesity and overweight prevalence: BMI
  - Many social welfare programs: from a normative perspective, perhaps even arguing for welfare weights
- Other reasons: median regression is more efficient if there's heteroskedasticity; easier to deal with censored data; not sensitive to outliers on the outcome variable

What is the n^{th} 100-quantile/percentile?

Suppose we sort a variable Y_i in ascending order, the n^{th} percentile of Y_i is the value that separates the first n percent of the values (ordering and sorting)

```
Y = rnorm(10000)
quantile(Y, probs= c(0.1, 0.5, 0.9))
```

```
        10%         50%         90% 
-1.27629928  0.01575219  1.27301209 
```

```
qnorm(c(0.025, 0.975))
```

```
[1] -1.959964  1.959964
```

What is τ-quantile?

- Suppose for an r.v. Y_i, we have a CDF $\mathrm{F}(y)$, which gives the probability $\Pr(\mathrm{Y}_i \leq y)$
- The quantile function of Y_i is the inverse CDF:

$$\mathrm{Q}_\tau(\mathrm{Y}_i) \equiv \mathrm{F}^{-1}(\tau)$$

  - The quantile function $\mathrm{Q}_\tau(\mathrm{Y}_i)$ returns the value y such that $\mathrm{F}(y) = \Pr(\mathrm{Y}_i \leq y)$

The Conditional Quantile Function

- The conditional quantile function at quantile τ given X_i is

$$\mathrm{Q}_\tau(\mathrm{Y}_i \mid \mathrm{X}_i) \equiv \mathrm{F}^{-1}(\tau \mid \mathrm{X}_i)$$

  - What is $\mathrm{Q}_{0.5}(\mathrm{Y}_i \mid \mathrm{X}_i)$? $\mathrm{Q}_{0.25}(\mathrm{Y}_i \mid \mathrm{X}_i)$?

Choice Under Uncertainty

- In microeconomics, we learned the expected utility framework. For example, a textbook example for deriving demand for insurance is

$$\max_x \mathrm{E}(u) = p \times u(y - d - qx + x) + (1 - p) \times u(y - qx)$$

- In general,

$$\mathrm{E}(u) = \int u(x) f(x) dx = \int u(x) d\mathrm{F}(x)$$

- Sometimes we further specify the functional form of $u(x)$, say Cobb-Douglas, quasi-linear, etc... And we often assume $u(x)$ is concave so that the solution exists

A Simple Statistical Decision Story

- Consider an econometrician observing some data $\mathrm{Y}_i, \mathrm{X}_i$, she want to make a choice to minimize some loss due to prediction errors in different states of the world
- Let the predicted error be $e_i = \mathrm{Y}_i - \hat{\mathrm{Y}}_i(\mathrm{X}_i)$, a loss function is $L(e_i) = L(\mathrm{Y}_i - \hat{\mathrm{Y}}_i(\mathrm{X}_i))$
- The econometrician's problem is

$$\min_{\hat{\mathrm{Y}}_i(\mathrm{X}_i)} \mathrm{E}(L(e_i))$$

- If the loss function is our familiar square error $L(e_i) = e_i^2$ (penalty is larger when the error is big), then the optimal $\hat{\mathrm{Y}}_i(\mathrm{X}_i)$ is the CEF $\mathrm{E}(\mathrm{Y}_i \mid \mathrm{X}_i)$ (MHE Theorem 3.1.2)

Other Loss Functions

- Turned out the CQFs, $Q_\tau(Y_i \mid X_i)$, are solutions for other loss functions:
  - Absolute error $L(e_i) = |e_i| \rightsquigarrow Q_{0.5}(Y_i \mid X_i)$
  - Asymmetric absolute error (the check function) $L(e_i) = 1(e_i > 0) \times \tau|e_i| + 1(e_i \leq 0) \times (1-\tau)|e_i| \rightsquigarrow Q_\tau(Y_i \mid X_i)$
  - The expected loss functions, or the risk function, are convex, solutions are easily achieved through linear programming
- The CQF is the decision function (or a strategy); other decision functions are said to be dominated

More on the Check Function

- Asymmetric absolute error loss function punishes our econometrician differently for over-prediction and under-prediction
    - Relevant example: Predicting flood levels; predicting distributional welfare effect ex ante; predicting demands for some perishable goods
- The check function

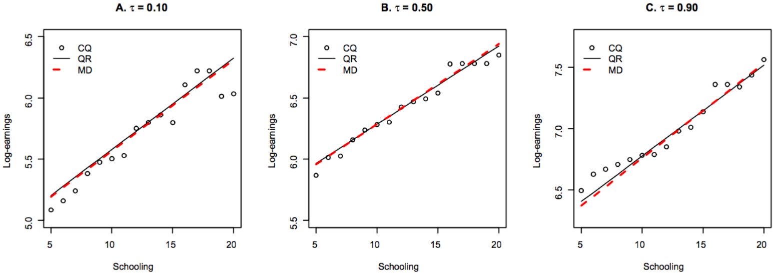
$$
\begin{aligned}
L(e_i) &= 1(e_i > 0) \times \tau|e_i| + 1(e_i \leq 0) \times (1-\tau)|e_i| \\
&= \begin{cases} \tau \times e_i & \text{if underpredicted} \\ (1-\tau) \times (-e_i) & \text{if overpredicted} \end{cases}
\end{aligned}
$$

- For higher percentile, the cost is higher for underprediction; for lower percentile, the cost is higher for overprediction
    - Specified quantiles deliver risk preference

Regression as Approximations

- Recall that we motivate running OLS regression as approximating the CEF
  - If CEF is linear, then OLS regression is it
  - If CEF is not linear, we still get a linear approximation
- When CQFs are of interest, quantile regressions approximate CQFs
  - MHE Theorem 7.1.1 *Quantile Regression Approximation*

Quantile Regression Approximation



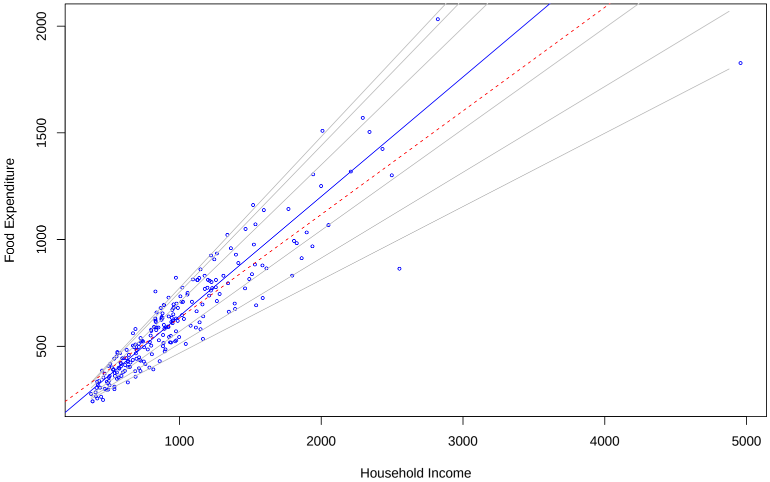
Quantile Engel Curves

- Engel's (1857) food expenditure data: 235 observations (working class household) on income and expenditure on food
- Quantile can be linked to this idea of "ordering and subsetting", so why not subsetting the outcome variable and then do OLS?

Quantile Engel Curves I

```
library(quantreg);data(engel);attach(engel)
plot(income,foodexp,xlab="Household Income",
     ylab="Food Expenditure",type = "n", cex=.5)
points(income,foodexp,cex=.5,col="blue")
taus <- c(.05,.1,.25,.75,.9,.95)
xx <- seq(min(income),max(income),100)
f <- coef(rq((foodexp)~(income),tau=taus))
yy <- cbind(1,xx)%*%f
for(i in 1:length(taus)){
        lines(xx,yy[,i],col = "gray")}
abline(lm(foodexp ~ income),col="red",lty = 2)
abline(rq(foodexp ~ income), col="blue")
```

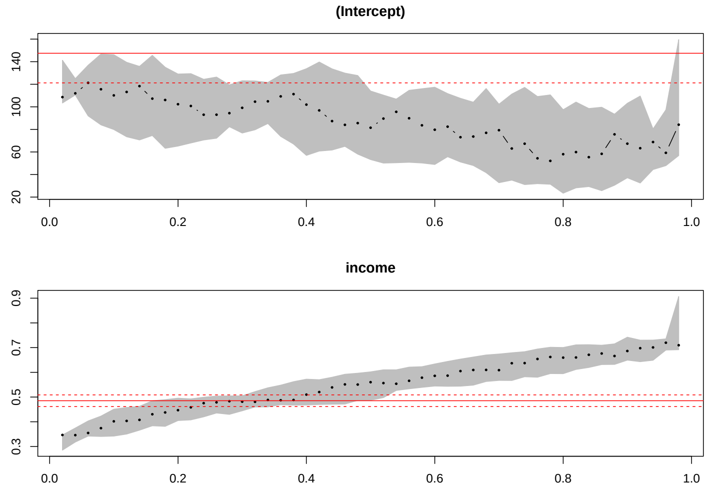
Quantile Engel Curves II



Quantile Engel Curves I

```
plot(summary(rq(foodexp~income,tau = 1:49/50,data=engel)))
```

Quantile Engel Curves II



Another Quantile Regression Plot I

```
medexp = foreign::read.dta(
    "http://cameron.econ.ucdavis.edu/musbook/mus03data.dta'
attach(medexp)
Y <- cbind(totexp)
X <- cbind(suppins, totchr, age, female, white)
quantreg <- rq(Y ~ X, tau = seq(0.05, 0.95, by = 0.05),
               data=medexp)
quantreg.plot <- summary(quantreg)
plot(quantreg.plot)
```

Another Quantile Regression Plot II