Machine Learning
Problem Set 2

Due Wednesday January 30th

Quantile regression is included in most modern statistical software packages. Implementing the estimator is as straightforward as running a single command, and the level of analysis far exceeds that of OLS. The goal of this problem set is to introduce you to quantile regression by having you apply it to analyze a data set.

Birth weight has a been found to be correlated with health outcomes during childhood and adulthood, as well as mental and physical development.[1] Low birth weight can thus be a real cause for concern. Its economic implications through health care and education costs may also be large. This exercise will let you carry out a brief analysis of how birth weight relates to various prenatal and demographic variables. Provided to you will be a natality data set for the US from 1996.[2] See Table 1 for a description of the variables.

Quantile regression is based on the following problem. Suppose you have a random variable Y. Define the loss function $\rho_\tau(y) = y(\tau - \mathbb{1}\{y < 0\})$, where $\tau \in [0, 1]$.[3] The solution $q^\star$ to the problem

$$\min_q \mathbb{E}\left[\rho_\tau\left(Y - q\right)\right] \tag{1}$$

satisfies $\tau = F_Y\left(q^\star\right)$. In other words, $q^\star$ is the τ^{th} quantile of Y.

1. Provide the intuition as to why the solution to (1) is the τ^{th} quantile of Y. Think about the shape of ρ_τ, where the loss is greatest, and how that depends on τ.

Conditional quantile regression extends the idea above to the case where we condition on $\mathbf{X} \in \mathbb{R}^k$. By assumption, the τ^{th} quantile function is $Q_{Y|\mathbf{X}}(\tau) = \mathbf{X}'\boldsymbol{\beta}_\tau$. The vector $\boldsymbol{\beta}_\tau$ can be recovered by solving the problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \mathbb{E}\left[\rho_\tau\left(Y - \mathbf{X}'\boldsymbol{\beta}\right)\right].$$

[1] This exercise is based on the paper by Abrevaya (2006).

[2] The data is obtained from here in case you would like to use it for your own research. The data sets are large. The data provided for this exercise are restricted to the complete observations of the first 10,000 observations from the 1996 sample.

[3] This is also called the 'check function', because of its similarity to a check mark.

Table 1: Variable names and descriptions

Variable name	**Description**
`birthweight`	Birth weight (g)
`boy`	Male indicator
`married`	Mother married
`black`	Mother black
`age`	Mother's age during pregnancy
`highschool`	Mother completed high school
`somecollege`	Mother completed some college
`college`	Mother completed college
`prenone`	No prenatal care
`presecond`	First received prenatal care in second trimester
`prethird`	First received prenatal care in third trimester
`cigsdaily`	Cigarettes smoked per day by mother
`weightgain`	Weight gain of mother during pregnancy (lbs)

The econometric model can be specified as follows:

$$\begin{aligned} Y &= Q_Y\left(U \mid \mathbf{X}\right) \\ U &\sim \mathit{Unif}\left[0,1\right] \\ Q_Y\left(\tau \mid \mathbf{X}\right) &= \mathbf{X}'\boldsymbol{\beta}_\tau \text{ for } \tau \in [0,1] \end{aligned} \tag{2}$$

Note that for each τ is a different $\boldsymbol{\beta}_\tau$.

2. What would be the interpretation of the coefficient estimates from running OLS of Y on $\mathbf{X}$? Likewise, what would be the interpretation of the coefficient estimates of a quantile regression (QR) of Y on $\mathbf{X}$, as in $\boldsymbol{\beta}_\tau$ in equation (2)? How do the two interpretations differ?

3. Suppose you wished to make a causal interpretation of the regression model. What assumptions are required for OLS? Will those assumptions differ for QR? If so, how?

4. Regress Y (`birthweight`) on $\mathbf{X}$ (all other variables, including `age^2` and `weightgain^2`) using OLS, and report your results in a nice LaTeXtable. Discuss any interesting relationships that you observe, and whether they make economic sense.

5. Now instead carry out the quantile regression for various levels of τ—the more values of τ you consider, the richer your set of results. Present the estimates for each coefficient in a nice set of figures, and discuss any interesting patterns that you observe. Do they make economic sense? This link may be helpful for R users.

6. Compare the OLS estimates against the QR estimates. How do the two relate? How well does OLS summarize the relationship between Y and $\mathbf{X}$ relative to QR?

7. Let $\widehat{\varepsilon}_i$ denote the residuals from your OLS regression in question 4. What is $\sum_{i=1}^{n} \widehat{\varepsilon}_i$? Also, how many of the residuals are 0? Provide an explanation for your findings.

8. Re-run your QR from question 5 for a τ of your choice. Let $\widetilde{\varepsilon}_i$ denote the residuals. What is $\sum_{i=1}^{n} \widetilde{\varepsilon}_i$? Also, how many of the residuals are (approximately) 0? You may have to tolerate some imprecision e.g.

$$h = \sum_{i=1}^{n} \mathbb{1}\left\{|\widetilde{\varepsilon}_i| < e^{-10}\right\}. \tag{3}$$

   How do your findings compare to those from question 7? Do your findings depend on τ?

9. Access the results of the dual problem (e.g. in R, if the results of your QR are stored in the object `results`, the solution to the dual problem can be accessed via `results$dual`). How many of these values are strictly between 0 and 1?

   Note: The importance of this finding, and its relation to your findings in question 8 will be explained later in the course.

10. Given your findings from question 8, can you think of a way to recover $\boldsymbol{\beta}_\tau$ for some τ using only using only h observations from the data set, where h is defined as in (3)?

    Hint: If you know p points on the fitted plane in $\mathbb{R}^p$, then you must be able to back out the fitted plane, as well as the regression coefficients that define it.

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics 21*(4), 489–519.