Problem Set 2

1.) Quantile Regression

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 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. You will be provided with a natality data set for the US from 1996.[2] See Table 1 for a description of the variables.

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)

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

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

[2] The original data is 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.

$q^\star$ 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.

Problem 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 $Q_\tau(Y \mid \mathbf{X}) = \mathbf{X}'\beta_\tau$. The vector β_τ can be recovered by solving the problem

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

The econometric model can be specified as follows:

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

Note that for each τ there is a different β_τ.

Problem 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 $\mathbf{X}$, such as that such β_τ in equation (2)? How do the two interpretations differ?*

Problem 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?*

Problem 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.*

Problem 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.*[4] *Do they make economic sense? Consider plotting β_τ against τ.*

Problem 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?*

[4]This link may be helpful for R users: `https://data.library.virginia.edu/getting-started-with-quantile-regression/`.

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

Problem 8. *Re-run your QR from Problem 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.*

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

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

Problem 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 Problem 8, are explained in the course. For now, you should just be intrigued.

Problem 10. *Given your findings from Problem 8, can you think of a way to recover β_τ 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.

2.) Linear Programming, from (Bertsimas and Tsitsiklis, 1997)

Problem 11. *Consider the problem*

$$\text{minimize } 2x_1 + 3|x_2 - 10|$$

$$\text{subject to } |x_1 + 2| + |x_2| \leq 5,$$

and reformulate it as a linear programming problem.

Problem 12. ***Bonus:*** *An investor has a portfolio of n different stocks. He has bought s_i shares of stock i at price p_i, $i = 1, \ldots, n$. The current price of one share of stock i is q_i. The investor expects that the price of one share of stock i in one year will be r_i. If he sells shares, the investor pays transaction costs at the rate of 1% of the amount transacted. In addition, the investor pays taxes at the rate of 30% on capital gains. For example, suppose that the investor sells 1,000 shares of a stock at \$50 per share. He has bought these shares at \$30 per share. He receives \$50,000. However, he owes 0.30×(50,000 − 30,000) = \$6,000 on capital*

gain taxes and 0.01 × (50,000) = \$500 on transaction costs. So, by selling 1,000 shares of this stock he nets 50,000 − 6,000 − 500 = \$43,500. Formulate the problem of selecting how many shares the investor needs to sell in order to raise an amount of money K, net of capital gains and transaction costs, while maximizing the expected value of his portfolio next year.

Problem 13. ***(Chebychev Center)*** *Consider a set P described by linear inequality constraints, that is, $P = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{a_i}'\mathbf{x} \leq b_i, i = 1, \ldots, m\}$. A ball with center $\mathbf{y}$ and radius r is defined as the set of all points within (Euclidean) distance r from $\mathbf{y}$. We are interested in finding a ball with the largest possible radius, which is entirely contained within the set P. (The center of such a ball is called the Chebychev center of P.) Provide a linear programming formulation of this problem. Additionally, put the problem in standard form.*

Problem 14. *Let $\mathbf{A}$ be a symmetric square matrix. Consider the linear programming problem*

$$\text{minimize } \mathbf{c}'\mathbf{x}$$

$$\text{subject to } \mathbf{A}\mathbf{x} \geq \mathbf{c},\ \mathbf{x} \geq 0.$$

Prove that if $\mathbf{x}^*$ satisfies $\mathbf{A}\mathbf{x}^* = \mathbf{c}$ and $\mathbf{x}^* \geq \mathbf{0}$, then $\mathbf{x}^*$ is an optimal solution.

Problem 15. *Show equivalence of the two dual formulations of the quantile regression problem:*

$$\begin{aligned} &\max_{\mathbf{d}} \ \mathbf{Y}^T\mathbf{d} \\ &\text{s.t. } \mathbf{X}^T\mathbf{d} = \mathbf{0} \\ &(\tau - 1)\mathbf{1}_n \leq \mathbf{d} \leq \tau\mathbf{1}_n \end{aligned} \tag{4}$$

and

$$\begin{aligned} &\max_{\mathbf{a}} \ \mathbf{Y}^T\mathbf{a} \\ &\text{s.t. } \mathbf{X}^T\mathbf{a} = (1 - \tau)\mathbf{X}^T\mathbf{1}_n \\ &\mathbf{a} \in [0, 1]^n. \end{aligned} \tag{5}$$

3.) A derivation of 2SLS via OVB formulas

Consider this unusual derivation of the two-stage least-squares estimator. We model the endogeneity as coming from an unobserved variable A measuring, say, underlying aptitude level. The regression function of interest is

$$E[Y|X, D, A] = X\beta_X + D\beta_D + A, \tag{6}$$

where X is exogenous and D is may be correlated with A, which is unobserved. We may think of D as measuring schooling, and X some measure of family background.

The first exclusion restriction is

$$E\left[Y\,|X, D, A, Z\right] = X\beta_X + D\beta_D + A. \tag{7}$$

In other words, the implicitly defined β_Z is 0. Equation (7) will serve a our "long regression".

The second exclusion restriction is

$$E^*\left[A\,|X, Z\right] = X\lambda_X. \tag{8}$$

This will hold under, for instance, random assignment of the instrument.

Problem 16. *Briefly explain the content of each of these restrictions.*

Next, consider the hypothetical setting of $\mathbb{E}[Y|U, V, W, R] = V\beta_V + W\beta_W + U\beta_U + R\beta_R$. Consider the regression modeling $\mathbb{E}^*[Y|V, W] = V\gamma_V + W\gamma_W$. Then the omitted variable bias formula says that

$$\begin{bmatrix}\gamma_V\\ \gamma_W\end{bmatrix} = \begin{bmatrix}\beta_V\\ \beta_W\end{bmatrix} + \mathbb{E}[((V,W)^T(V,W))^{-1}]\mathbb{E}[(V,W)^T(U,R)] \begin{bmatrix}\beta_U\\ \beta_R\end{bmatrix}.$$

Problem 17. *Returning to our setting, Our "short regression" will be*

$$E^*\left[Y\,|X, Z\right] = X\gamma_X + Z\gamma_Z. \tag{9}$$

Use the omitted variable bias formula to produce an expression for $\begin{bmatrix}\gamma_X\\ \gamma_Z\end{bmatrix}$.

Now, define $\gamma_{D\sim X+Z} \in \mathbb{R}^{(p_X+p_Z)\times p_D}$ is the best linear predictor coefficient for the regression of D on (X, Z), and $\gamma_{A\sim X+Z}$ is defined analogously.

Problem 18. *From your answer to Problem 17, derive*

$$\beta_D = \left(\gamma_{D\sim X+Z}\right)_Z^{-1}\gamma_Z, \tag{10}$$

where $\left(\gamma_{D\sim X+Z}\right)_Z$ *corresponds to the* p_Z *lower rows of* $\gamma_{D\sim X+Z}$*. This is the two-stage least-squares estimator!*

Problem 19. *Consider for simplicity the case without covariates* X*. From the formula above, show that*

$$\beta_D = E\left[Z^T D\right]^{-1} E\left[Z^T Y\right].$$

References

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

Bertsimas, Dimitris, and John N Tsitsiklis. 1997. *Introduction to linear optimization.* Vol. 6, Athena Scientific Belmont, MA.