Introduction To Quantile Regression

Hanxiang Peng

Department of Mathematical Sciences
Indiana University-Purdue University at Indianapolis

August 30, 2010

Outline

Linear-Regression Modeling and Its Shortcomings

- y=household income. x=interval variable, ED (the household head's years of schooling), or a dummy variable, BLACK (the head's race, 1 = black and 0 = white). Data: $(x_i, y_i) : i = 1, ..., n$.
- In linear regression model (LRM),

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

  Assumptions: ϵ_i iid N(0, σ^2).
- $E(y|x) = \beta_0 + \beta_1 x$: the average in the population of y values corresponding to a fixed value of the covariate x.

- $\hat{y} = 23127 + 5633ED$

ED	9	12	16
$E(y\|ED)$	\$27,570	\$44,469	\$67,001

- $\hat{y} = 53466 - 18268BLACK$

BLACK	0	1
$E(y\|BLACK)$	\$53,466	\$35,198

- In LRM: **the mean of a distribution representing its central tendency; homoscedasticity assumption; normality assumption; Outliers** (The usual practice is to identify outliers and eliminate them. Both the notion of outliers and the practice of eliminating outliers undermine much social-science research, particularly studies on social stratification and inequality, as outliers and their relative positions to those of the majority are important aspects of inquiry).

Household Income Data

- The location shifts among 3 education groups and between blacks and whites are obvious, their shape shifts are substantial.
- The conditional mean from the LRM fails to capture the shape shifts caused by changes in the covariate (education or race)
- Since the spreads differ substantially among the education groups and between the two racial groups, the **homoscedasticity** assumption is violated.
- All box graphs are **right-skewed**. LRM models are not able to detect these shape changes.
- Seven **outliers** identified: three cases with 18 years of schooling having an income of more than $ 505,215 and four cases with 20 years of schooling having an income of more than $471,572.

INCOME INEQUALITY IN 1991 AND 2001

Quantiles

- The pth quantile $Q^{(p)}$ of a cdf F is the minimum of the set of values y such that $F(y) \geq p$ for $0 \leq p \leq 1$. The function $Q^{(p)}$ (as a function of p) is referred to as the quantile function.
- Given a sample $y_1, ..., y_n$, the pth sample quantile $\hat{Q}^{(p)}$ is the pth quantile of the corresponding empirical cdf $\hat{F}$; $\hat{Q}^{(p)}$ is the sample quantile function.

Quantile as solution to minimization problem

- (Sample) mean as solution to minimization: $\hat{\mu} = \bar{y}$ solves

$$\min_{\mu} \sum (y_i - \mu)^2$$

- Median as solution to min: $\hat{m} = median(y_1, ..., y_n)$ solves

$$\min_{m} \sum |y_i - m|$$

- Quantile as solution to min: $\hat{Q}^{(p)}$ solves

$$\min_{q} \left\{ (1-p) \sum_{y_i < q} |y_i - q| + p \sum_{y_i \geq q} |y_i - q| \right\}$$

- **Monotone equivariance**: Suppose h is monotone. If q is the pth quantile of Y, then $h(q)$ is the pth quantile of $h(Y)$.
- **Robust to outliers**.

Quantile regression models (QRM)

- In LRM,

$$E(y|x) = \beta_0 + \beta_1 x$$

- In QRM, for $0 < p < 1$,

$$Q^{(p)}(y|x) = \beta_0^{(p)} + \beta_1^{(p)} x, \quad 0 < p < 1$$

Example: Fitting Household Income Data

- Tables 3.2 and 3.3: 19 conditional quantiles of income given education or race; the coefficient for education grows monotonically from $1,019 at the .05th quantile to $ 8,385 at the .95th quantile. Similarly, the black effect is weaker at the lower quantiles than at the higher quantiles.
- Conditional quantiles on 12 years of schooling:

p	.05	.50	.95
$\hat{Q}^{(p)}(y_i \| ED_i = 12)$	$7,976	$36,727	$111,268

- Conditional quantiles on blacks:

p	.05	.50	.95
$\hat{Q}^{(p)}(y_i \| BLACK_i = 1)$	$5,432	$26,764	$91,761

- These results are very different from the conditional mean of the LRM.

- The left panel of Figure 3.3 presents the scatterplot of household income against the head of household's years of schooling. The single regression line indicates mean shifts, for example, a mean shift of $22,532 from 12 years of schooling to 16 years of schooling ($5633 \cdot (16 - 12)$). However, this regression line does not capture shape shifts.
- The right panel of Figure 3.3 shows the same scatterplot as in the left panel and the 19 quantile-regression lines. The .5th quantile (the median) fit captures the central location shifts, indicating a positive relationship between conditional-median income and education. The slope is $4,208, shifting $16,832 from 12 years of schooling to 16 years of schooling ($4208 \cdot (16 - 12)$). This shift is lower than the LRM mean shift.

- In addition to the estimated location shifts, the other 18 quantile-regression lines provide information about shape shifts. These regression lines are positive, but with different slopes. The regression lines **cluster tightly a low levels of education** (e.g., 0-5 years of schooling) but **deviate from each other more widely at higher levels of education** (e.g., 16-20 years of schooling).
- A shape shift is described by the tight cluster of the slopes at lower levels of education and the scattering of slopes at higher levels of education. For example: the spread of the conditional income on 16 years of schooling (from $12,052 for the .05th conditional quantile to $144,808 for the .95th conditional quantile) is much wider than that on 12 years of schooling (from $7,976 for the .05th conditional quantile to $111,268 for the .95th conditional quantile).

QR Estimation

- In LRM, the least squres estimates $\hat{\beta}_1, \hat{\beta}_2$ solves

$$\min_{\beta_1,\beta_2} \sum (y_i - \beta_1 - \beta_2 x_i)^2$$

- In median-regression model ($p = .5$), the estimates $\hat{\beta}_1^{(.5)}, \hat{\beta}_2^{(.5)}$ solves

$$\min_{\beta_1,\beta_2} \sum |y_i - \beta_1 - \beta_2 x_i|$$

  the resulting median-regression line, must pass through a pair of data points with half of the remaining data lying above the regression line and the other half falling below.

QR Estimation

- In QRM ($0 < p < 1$), the estimates $\hat{\beta}_1^{(p)}, \hat{\beta}_2^{(p)}$ solves

$$\min_{\beta_1,\beta_2} \left\{ (1-p) \sum_{y_i < \beta_1^{(p)} + \beta_2^{(p)} x_i} |y_i - \beta_1 - \beta_2 x_i| + p \sum_{y_i \geq \beta_1^{(p)} + \beta_2^{(p)} x_i} |y_i - \beta_1 - \beta_2 x_i| \right\}$$

  the resulting pth quantile regression estimator must pass through a pair of data points with p proportion of data points lying below the fitted line $y = \hat{\beta}_1^{(p)} + \hat{\beta}_2^{(p)} x$, and the $1 - p$ proportion lying above.
- This is a linear programming problem and algorithms for computing the quantile-regression coefficients have been developed.

- For example, when we estimate the coefficients for the .10th quantile regression line, the observations below the line are given a weight of .90 and the ones above the line receive a smaller weight of .10. As a result, 90% of the data points (x_i, y_i) lie above the fitted line leading to positive residuals, and 10% lie below the line and thus have negative residuals.
- Conversely, to estimate the coefficients for the .90th quantile regression, points below the line are given a weight of .10, and the rest have a weight of .90; as a result, 90% of observations have negative residuals and the remaining 10% have positive residuals.

Transformation, Equivariance and Robustness

- In LRM,
  $$E(c + ay|x) = c + aE(y|x)$$
  Similar for QRM: $a > 0$ or $a < 0$
  $$Q^{(p)}(c{+}ay|x) = c{+}aQ^{(p)}(y|x) \; or \; Q^{(p)}(c{+}ay|x) = c{+}aQ^{(1-p)}(y|x)$$
- Monotone equivariance: if h is monotone (incr), then
  $$Q^{(p)}(h(y)|x) = h(Q^{(p)}(y|x))$$
  LRM does not have this property.
- Robustness: the QRM estimates are not sensitive to outliers. LRM is not robust.

Conditions

Assume that $Z_1, \ldots, Z_n$ are independent replicates of $Z = (X^\top, Y)^\top$ which form the linear regression model

$$Y = \beta^\top X + \varepsilon, \qquad (1)$$

where β is a parameter, $E(XX^\top)$ is finite and positive definite, and ε is an unobservable random error that has continuous conditional density $f(t|X)$ given X, bounded and bounded away from zero at $t = 0$ and satisfying $\int_{-\infty}^{0} f(t|X)dt = p$ for $0 < p < 1$ and $E(f(0|X)XX^\top)$ positive definite. The quantile regression estimator $\hat{\beta}^{(p)}$ of β solves:

$$\hat{\beta}^{(p)} = \arg\min_b \sum_j \rho_p(Y_j - b^\top X_j), \qquad (2)$$

where $\rho_p(t) = (p - \mathbf{1}[t < 0])t, t \in \mathbb{R}$ is the check function.

Theorem

Under the above conditions, $\hat{\beta}^{(p)}$ has an asymptotic normal distribution with mean β_0 and variance-covariance matrix

$$\frac{p(1-p)}{nf^2(0|X)}\left\{E(XX^\top)\right\}^{-1}.$$

-
-