ถวิล นิลใบ
เศรษฐศาสตร์ รามคำแหง

ควอนไทล์รีเกรสชั่น (Quantile Regression)[1]

เนื้อหาของเอกสารแบ่งเป็น 5 ส่วน ส่วนแรก กล่าวถึง หลักการและพัฒนาการของควอนไทล์รีเกรสชั่น ส่วนที่สอง กล่าวถึง ข้อจำกัดของการวิเคราะห์สมการถดถอยแบบดั้งเดิมที่ใช้กันแพร่หลาย หรือที่เรียกว่า "Mean Regression" ส่วนที่สาม กล่าวถึง ความหมายของคำว่า "quantile" "Quantile Functions" และ "Quantile Regression Functions" ส่วนที่สี่ วิธีการคำนวณควอนไทล์รีเกรสชั่นและการประเมิน และส่วนสุดท้าย นำเสนอตัวอย่างการประยุกต์ใช้ควอนไทล์รีเกรสชั่น

1. หลักการและพัฒนาการของควอนไทล์รีเกรสชั่น

วัตถุประสงค์ของการวิเคราะห์สมการถดถอย (regression analysis) คือการหาความสัมพันธ์ระหว่างตัวแปรที่สนองตอบ (a response variable หรือตัวแปรตาม) ต่อตัวแปรกำหนด (predictor variables หรือตัวแปรอิสระ) อย่างไรก็ตาม ความสัมพันธ์ของตัวแปรทั้งสองไม่ได้มีลักษณะที่แน่นอน (stochastic relationship) หรือเรียกว่า ค่าตัวแปรสนองตอบมีลักษณะตัวแปรเชิงสุ่ม (a random variable) กล่าวคือ ถ้าค่าของตัวแปรกำหนดแต่ละค่า จะให้ค่าของตัวแปรสนองตอบได้หลาย ๆ ค่า ด้วยความน่าจะเป็นที่แตกต่างกัน ซึ่งเรียกโดยสรุปว่า จะมีการแจกแจงความน่าจะเป็นของค่าของตัวแปรตามหรือตัวแปรสนองตอบ ณ แต่ละค่าของตัวแปรกำหนดหรือตัวแปรอิสระ ดังนั้น การวิเคราะห์ของตัวแปรสนองตอบต่อตัวแปรกำหนดจึงใช้วิธีการวัดแน้วโน้มเข้าสู่ส่วนกลาง (measures of central tendency) เพื่อเป็นตัวแทนผลของการสนองตอบ การวิเคราะห์สมการถดถอยรูปแบบดั้งเดิมจึงเป็นการวัดแนวโน้มเข้าสู่ส่วนกลาง (central tendency) และนิยมใช้ค่าเฉลี่ย (mean) หรือค่าที่คาดหวัง (expectation) เป็นตัวแทนหลัก การวิเคราะห์สมการถดถอยที่ใช้ค่าเฉลี่ย (เรียกว่า "conditional mean regression" หรือ เรียกสั้นๆ ว่า "mean regression") มีข้อกำหนดหลายๆ ประการ ซึ่งข้อกำหนดดังกล่าวบ่อยครั้งมักไม่เป็นจริง เช่น การแจกแจงของตัวแปรตามที่สนองตอบต่อตัวแปรกำหนดต้องมีลักษณะสมมาตร (normal distribution) หรือข้อกำหนดว่า ความแปรปรวนของตัวรบกวนต้องมีค่าคงที่ (homoscedasticity) แต่ความเป็นจริงการแจกแจงอาจจะมีความเบ้ (skewed) รวมทั้งความแปรปรวนของตัวรบกวนมีค่าไม่คงที่ (heteroscedasticity) เป็นต้น นอกจากนี้ การวิเคราะห์สมการ

[1] เอกสารฉบับนี้เขียนขึ้นเพื่อประกอบการบรรยายให้กับคณาจารย์คณะเศรษฐศาสตร์ มหาวิทยาลัยรามคำแหง วันที่ 27 กุมภาพันธ์ 2562 ซึ่งมีหลายส่วนที่ยังไม่สมบูรณ์ ผู้เขียนยินดีรับฟังข้อเสนอแนะเพื่อนำไปปรับปรุง

ถดถอยที่ใช้ค่าเฉลี่ย ไม่สามารถขยายให้ครอบคลุมไปส่วนอื่นๆ ของการแจกแจงของค่าตัวแปรสนองตอบ (คือที่ปลายหางของการแจกแจง) ทำให้การวิเคราะห์ไม่สามารถตอบคำถามของการสนองตอบของกลุ่มที่มากกว่าค่าเฉลี่ย หรือ ต่ำกว่าค่าเฉลี่ย ตัวอย่างเช่น ในการศึกษาเรื่องการกระจายรายได้ นักวิจัยมักสนใจกลุ่มที่มีรายได้น้อย ซึ่งเป็นกลุ่มที่มีสัดส่วนสูง หรือการศึกษาผลสัมฤทธิ์ของการศึกษา นักวิจัยจะสนใจกลุ่มผู้มีผลสัมฤทธิ์ทางการศึกษาที่อยู่ในระดับต่ำ (ทั้งสองกลุ่มนี้ อยู่ปลายหางของการแจกแจง) เป็นต้น ซึ่งการวิเคราะห์จากสมการถดถอยแบบค่าเฉลี่ย จะไม่สามารถตอบคำถามเหล่านี้ได้ การวิเคราะห์ Quantile regression (QR) จึงเป็นทางเลือกในกรณีที่การแจกแจงมีลักษณะไม่สมมาตร หรือใช้ร่วมกับการวิเคราะห์สมการถดถอยแบบค่าเฉลี่ย (mean regression: MR) ซึ่งจะทำให้การวิเคราะห์ได้ครอบคลุมมากยิ่งขึ้น ทั้งนี้เพราะ การด้วยวิธีควอนไทล์รีเกรสชั่น (QR) นั้นหาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามที่ไม่ได้พิจารณาแต่เฉพาะแนวโน้มเข้าสู่ค่าเฉลี่ย แต่สามารถหาความสัมพันธ์ (สมการถดถอย) ณ จุดใดจุดหนึ่งของการแจกแจง หรือที่ ณ ระดับควอนไทล์ต่างๆ กัน

Quantile regression พัฒนาโดย Koenker และ Bassett (1978)[2] หลังจากนั้น ได้มีการนำQR มาประยุกต์ใช้ในหลายๆ ศาสตร์ เพื่อปรับปรุงการวิเคราะห์สมการถดถอยแบบค่าเฉลี่ย (MR) ในยุคเริ่มต้นของการพัฒนาตัวแบบ QR ยังคงวัดเข้าสู่ส่วนกลาง แต่ใช้ จุดมัธยฐาน (median) เรียกว่าตัวแบบ "median-regression model" หลักการคำนวณค่าพารามิเตอร์ของตัวแบบมัธยฐาน คือหาค่าพารามิเตอร์ที่ทำให้ " ระยะห่างระหว่างค่าคำนวณกับค่าสังเกตมีค่าต่ำที่สุด โดยไม่พิจารณาเครื่องหมาย (least-absolute distance estimation)" แต่การคำนวณหาค่าพารามิเตอร์ด้วยวิธี least-absolute distance ค่อนข้างยุ่งยาก เมื่อเทียบกับการคำนวณหาค่าพารามิเตอร์ของตัวแบบ mean regression ที่คำนวณด้วยวิธี Least squares ซึ่งไม่ยุ่งยากและไม่ต้องใช้เวลามาก การคำนวณตัวแบบ median regression ต้องใช้เครื่องคอมพิวเตอร์ที่มีประสิทธิภาพสูงในการคำนวณ อย่างไรก็ตาม ปัจจุบัน ได้มีการพัฒนาประสิทธิภาพการคำนวณของคอมพิวเตอร์ ทำให้สามารถคำนวณด้วยวิธีนี้ได้ไม่ยากเหมือนเมื่อก่อน การสร้างตัวแบบ median-regression model ซึ่งเป็นตัวแบบหนึ่งของ Quantile regression สามารถนำมาใช้เป็นทางเลือกของ mean-regression model เพื่อแสดงความสัมพันธ์ระหว่างค่าแนวโน้มเข้าสู่ส่วนกลางของตัวแปรตามที่จะสนองตอบต่อกลุ่มของตัวแปรอิสระที่เปลี่ยนไป ต่อมา QR ได้รับความนิยมเพิ่มมากขึ้น ทั้งนี้เพราะประสิทธิภาพของเครื่องคอมพิวเตอร์เพิ่มมากขึ้น จึงช่วยทำให้การคำนวณรวดเร็วและง่ายขึ้น รวมทั้งมีโปรแกรมสำเร็จรูปช่วยคำนวณ เช่น STATA และ EViews เป็นต้น ประกอบกับวิธี QR ยังมีจุดเด่นเมื่อเปรียบเทียบกับวิธีการคำนวณถดถอยแบบดั้งเดิม (Mean regression: MR) ในหลายประเด็นที่สำคัญ เช่น ความแปรปรวนของตัวรบกวนที่เกิดขึ้นในแต่ละค่าของตัวกำหนดไม่เท่ากัน (heteroscedasticity) หรือกรณีที่การแจกแจงที่มีลักษณะไม่สมมาตรหรือมีความเบ้ (skewed) มากๆ QR จะมีความยืดหยุ่นมากกว่า

[2] Koenker, R. and Bassett, G. (1978) Regression Quantiles. Econometrica, 46, 33-50. http://dx.doi.org/10.2307/1913643

วิธีการวิเคราะห์แบบ MR ที่มุ่งเฉพาะเจาะจงไปที่ค่าเฉลี่ย แต่ QR สามารถเปลี่ยนที่ตั้งของการแจกแจงของผลลัพท์ได้ว่าจะใช้ที่ตำแหน่งใด (quantile ที่กำหนด) เช่น การแจกแจงที่ปลายหางด้านล่าง (a location at the lower tail) หรือที่ quantile ที่ 0.1 หรือ การแจกแจงที่ปลายหางด้านบน (a location at the upper tail) หรือที่ quantile ที่ 0.9 เป็นต้น

2. ข้อจำกัดของการวิเคราะห์สมการถดถอยแบบ " Mean Regression (MR)"

การแสดงความสัมพันธ์ระหว่างตัวแปรสนองตอบกับ ณ แต่ละค่าของตัวแปรกำหนดด้วยฟังก์ชั่นที่เรียกว่า "conditional mean of the response" ตัวแบบสมการถดถอยแบบ Conditional-mean models (MR) มีจุดเด่นหรือคุณสมบัติที่ทำให้วิธีนี้ได้รับความนิยม กล่าวคือ การคำนวณหาค่าพารามิเตอร์ทำได้ง่าย รวดเร็วเช่น ด้วยวิธี OLS หรือ วิธี maximum likelihood และให้ตัวคำนวณที่มีคุณสมบัติ และง่ายต่อการตีความหมาย แต่อย่างไรก็ตาม การวิเคราะห์สมการถดถอยแบบ MR มีข้อจำกัด หลายๆ ประการดังต่อไปนี้

1. การวิเคราะห์ภายใต้ตัวแบบ MR ไม่สามารถที่จะขยายการวิเคราะห์การสนองตอบของตัวแปรตามต่อปัจจัยกำหนดให้ครอบคลุมถึงจุดอื่นๆ ของการแจกแจงของค่าสนองตอบของตัวแปรตาม (noncentral locations) นอกเหนือจุดค่าเฉลี่ย (mean) เช่น ค่าที่ปลายหางทั้งสองข้าง (lower and upper tails) ของการแจกแจง ค่าเหล่านี้อยู่นอกเหนือขอบเขตการวิเคราะห์ (outliers) ค่าเหล่านี้บางกรณีก็ให้ความรู้ที่เป็นประโยชน์กับนักวิจัย แต่ถูกละเลยในการวิเคราะห์ ยกตัวอย่างเช่น การศึกษาเกี่ยวกับความไม่เท่าเทียมทางด้านเศรษฐกิจ เช่น ทางด้านรายได้ หรือ ด้านค่าจ้าง หรือ ด้านการศึกษา นักวิจัยอาจจะสนใจวิเคราะห์กลุ่มผู้มีรายได้น้อยหรือกลุ่มคนจน (ซึ่งอยู่ปลายหางด้านล่างของการแจกแจง) และกลุ่มคนรวย (อยู่ปลายหางด้านบน). หรือการวิเคราะห์กลุ่มที่ได้รับค่าจ้างสูงหรือต่ำกว่าระดับเฉลี่ย เป็นต้น

2. ข้อกำหนดหรือสมมุติฐาน (assumptions) ที่อยู่เบื้องหลังการวิเคราะห์ของตัวแบบ MR หลายสมมุติฐานมักจะไม่เกิดขึ้นจริง โดยเฉพาะ สมมุติฐานเรื่อง ความแปรปรวนของตัวรบกวนมีค่าคงที่ (homoscedasticity) กรณีความแปรปรวนไม่คงที่ (เรียกกรณีนี้ว่า "scale shift") เช่น มีค่าเพิ่มขึ้น แสดงว่าเมื่อค่าของตัวแปรอิสระสูงขึ้น การแจกแจงของค่าสนองตอบหรือค่าของตัวแปรตามจะมีการกระจายมากขึ้น (ดูภาพที่ 1 ประกอบ) ซึ่งวัดจากค่าความแปรปรวน (variance) เมื่อค่าของตัวแปรสนองตอบจะกระจายห่างจากจุดค่าเฉลี่ยมาก กล่าวอีกนัยหนึ่งค่าสนองตอบที่อยู่ปลายหางทั้งสองข้างของการแจกแจงยิ่งมีมาก ซึ่งเป็นค่าที่อยู่นอกเหนือการวิเคราะห์ของสมการถดถอยแบบค่าเฉลี่ย ซึ่งการวิเคราะห์ที่มุ่งไปที่แนวโน้มค่าเฉลี่ย จะทำให้ผลการคำนวณไม่น่าเชื่อถือ แม้ว่า การวิเคราะห์แบบ MR จะแก้ปัญหาความแปรปรวนไม่คงที่ด้วยวิธี weighted least squares แต่ก็ยังคงวัดแนวโน้มเข้าสู่ค่าเฉลี่ย (mean)

3. กรณีการแจกแจงมีความแบ้ (skewness) เป็นกรณีที่การแจกแจงของค่าสนองตอบหรือค่าของตัวแปรตามมีความเบ้ (เรียกกรณีนี้ว่า shape shift) ซึ่งอาจจะเบ้ทางซ้ายหรือขวา (ดูภาพที่ 2 ประกอบ) จะทำค่าของตัวแปรสนองตอบกระจุกตัวอยู่ปลายหางมาก จะทำให้การวิเคราะห์ที่แนวโน้มค่าเฉลี่ยอาจจะไม่เหมาะสม ทั้งนี้เพราะค่าที่อยู่นอกเหนือของเขตความสนใจ (ห่างจากค่าเฉลี่ยมากๆ) จะมีอิทธิพลหรือบทบาทมาก แต่ถูกละเลยไม่ได้นำมาวิเคราะห์

4. การวิเคราะห์ที่จุดศูนย์กลางของการแจกแจง ทำให้นักวิจัยไม่ได้สนใจคุณสมบัติทั้งหมดของการแจกแจง (the whole distribution) การวิเคราะห์ควรกระทำมากกว่าการวิเคราะห์ตำแหน่งหรือจุดเดียวของการแจกแจง การวิเคราะห์ควรตั้งคำถามว่า การเปลี่ยนแปลงของตัวแปรอิสระจะมีจะมีผลอย่างไรต่อการเปลี่ยนแปลงต่อลักษณะของการแจกแจงของตัวแปรตาม ซึ่งจะทำให้การวิเคราะห์ได้ละเอียดลึกซึ้งและเป็นประโยชน์มากกว่า งานวิจัยทางด้านสังคมศาสตร์หลายๆ งานวิจัยมุ่งเน้นการวิเคราะห์การแบ่งชนชั้นทางสังคม (social stratification) ความเหลื่อมล้ำ ซึ่งเป็นประเด็นที่จะต้องมีการวิเคราะห์ในรายละเอียดที่เชื่อมโยงกับคุณสมบัติหรือคุณลักษณะของการแจกแจง ยกตัวอย่างเช่น การศึกษาเรื่องความไม่เท่าเทียมทางด้านค่าจ้าง รายได้และสินทรัพย์ หรือ ความไม่เท่าเทียมทางด้านการศึกษา หรือทางด้านสาธารณสุข เป็นต้น การวิเคราะห์ประเด็นเหล่านี้ด้วยวิธีแนวโน้มเข้าสู่ส่วนกลางจะไม่สามารถอธิบายได้อย่างชัดเจน

เปรียบเทียบความแตกต่างระหว่างการวิเคราะห์แบบ MR และ QR

Mean Regression (MR)	**Quantile Regression (QR)**
1. วิเคราะห์ที่จุดเดียวคือที่ค่าเฉลี่ย	1. วิเคราะห์ได้ทุกจุดของการแจกแจง
2. มีข้อกำหนดหรือสมมุติฐานที่เข้มงวด เช่น การแจกแจงแบบปกติที่มีลักษณะสมมาตร ความแปรปรวนมีค่าคงที่	2. มีข้อกำหนดหรือสมมุติฐานไม่เข้มงวด และสามารถจัดการกับการแจกแจงแบบที่ไม่มีลักษณะสมมาตร คือมีความเบ้ (skewness) หรือความแปรปรวนไม่คงที่
3. ไม่สามารถแก้ปัญหาเรื่อง (location shift) ความแปรปรวนไม่คงที่ (scale shift) ความเบ้ (shape shift)	3. แก้ปัญหาเรื่อง location shift, scale shift, shape shift

QR จะไม่พิจารณาตัวอย่างหรือประชากรทั้งหมดแล้วหาค่าเฉลี่ย แต่คำนึงถึงความแตกต่างระหว่างกลุ่มตัวอย่างหรือกลุ่มประชากรที่พิจารณา ซึ่งการสนองตอบต่อปัจจัยกำหนดจะแตกต่างกันระหว่างกลุ่ม (heterogeneous) กล่าวอีกนัยคือ การคำนวณหาสมการถดถอย ณ ระดับควอนไทล์ต่างๆ กันแล้วนำมาเปรียบเทียบกัน (ค่าพารามิเตอร์จะแตกต่างกันระหว่างกลุ่มหรือระหว่างแต่ละควอนไตล์)

3. ความหมายของคำว่า "quantile" Quantile Functions และ Quantile Regression Functions

ควอนไตล์ที่ p (pth quantile) แสดงถึง ค่าของของตัวแปรสนองตอบหรือตัวแปรตาม ซึ่งอยู่ต่ำกว่าค่าสนองตอบของประชากรทั้งหมดหรือการแจกแจงทั้งหมด (entire population) คิดเป็นสัดส่วนที่ p ควอนไตล์ที่ p ของค่า y มีค่าเท่ากับ μ_p เขียนเป็นสมการคือ

$$P(y \leq \mu_p) = F_y(\mu_p) = p \quad \text{..... (1)}$$

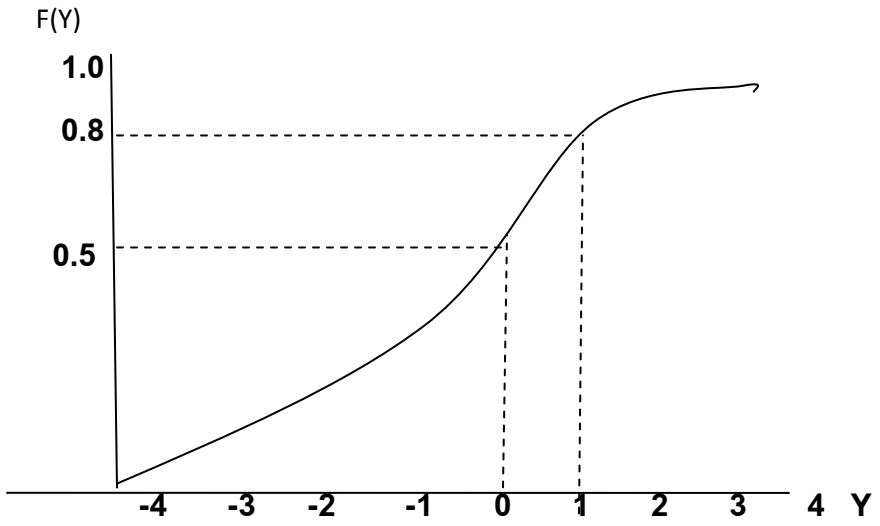
ตัวอย่างเช่น $Pr(y \leq 0) = F_Y(0) = 0.5$ ซึ่งหมายความว่า ความน่าจะเป็นที่ค่า Y จะมีค่าน้อยกว่าหรือเท่ากับ 0 มีค่าเท่ากับ 0.5 (ดูจากภาพที่ 3) หรือกรณีที่ $Pr(y \leq 1) = F_Y(1) = 0.8$ หมายความว่า ความน่าจะเป็นที่ค่า Y จะมีค่าน้อยกว่าหรือเท่ากับ 1 มีค่าเทากับ 0.8 เป็นต้น กรณีที่เราจะหาค่าความน่าจะเป็นที่ Y จะมีค่ามากกว่าค่าที่กำหนดจะเท่ากับ P(Y>y) = 1-F(y)

ดังนั้น เราสามารถกำหนดค่าของควอนไตล์ณ จุดใดจุดหนึ่งของการแจกแจง ยกตัวอย่างเช่น กำหนดควอนไตล์ที่ 0.5 หรือเขียนเป็นสัญญลักษณ์ $Q^{0.5} = 0$หมายถึงมีจำนวน 50% ของประชากรอยู่ต่ำกว่าค่า 0 หรือกำหนด กำหนดควอนไตล์ที่ 0.8 หรือเขียนเป็นสัญญลักษณ์ $Q^{0.8} = 1$ หมายถึงมีจำนวน 80% ของประชากร (ค่า Y ทั้งหมด) อยู่ต่ำกว่าค่า 1 เป็นต้น ดังนั้น $\mu_p = F_y^{-1}(p)$

Quantile Regression Functions

จากภาพที่ 4 มีสมการถดถอย 3 เส้น หรือ 3 สมการ สมการแรก $E(Y^m)$ เป็นสมการถดถอยที่คำนวณผ่านจุดค่าเฉลี่ยของการแจกแจงของความจะเป็นของตัวแปรตาม Y คือ E(Y) = a + bX หรือเขียนในรูปทั่วไป คือ

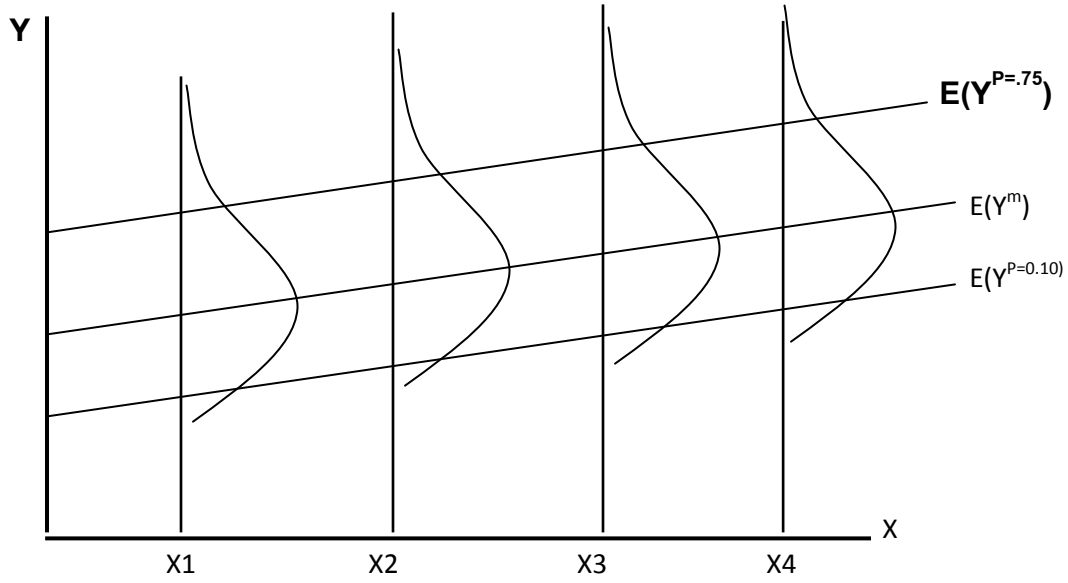
ภาพที่ 3: 1 การแจกแจงความน่าจะเป็นสะสม (Cummulative density function) ของการแจกแจงแบบปกติ



$$E(Y/X) = X\beta_m$$

ภาพที่ 4: สมการถดถอยที่ลากผ่านจุดเฉลี่ยและควอนไตล์ที่ 1 และ 3

กรณี "Homogeneous regression model"



สำหรับอีก 2 สมการ คือสมการเป็นสมการถดถอยที่คำนวณผ่านจุดค่าควอนไทล์ที่ 0.10 และ 0.75 ของการแจกแจงของความจะเป็นของตัวแปรตาม Y ได้แก่สมการ $E(Y^{P=0.10}) = a + bX$ และ $E(Y^{P=0.75}) = a + bX$ ตามลำดับ เขียนในรูปสมการทั่วไปคือ

$$E(Y/X) = X\beta_{0.10} \qquad \text{และ} \qquad E(Y/X) = X\beta_{0.75}$$

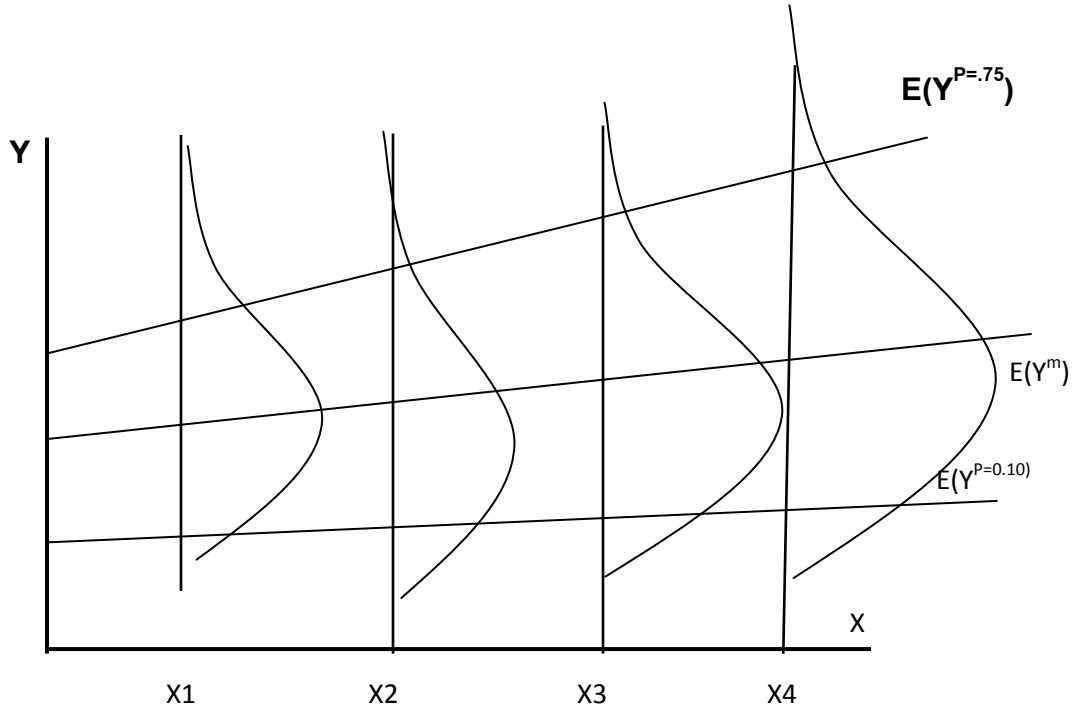
สมการถดถอยแบบควอนไทล์สามารถสร้างสมการถดถอยที่คำนวณขึ้น ณ จุดใดจุดหนึ่งของการแจกแจงของความน่าจะเป็นของตัวแปรตามได้ ไม่เจาะจงไปที่จุดค่าเฉลี่ย ดังนั้น ผลการวิเคราะห์สมการถดถอยแบบควอนไทล์จึงวิเคราะห์ในรายละเอียดของค่าตัวแปรตามได้ในทุก ๆ สถานการณ์ ซึ่งการวิเคราะห์ด้วยสมการถดถอยแบบดั้งเดิมมีข้อบกพร่อง ได้แก่ กรณี การแจกแจงที่ค่าของปลายหางของการแจกแจงถ้าลักษณะการแจกแจงของตัวแปรตามมีความเบ้ (skewness)[3] คือไม่สมมาตร (asymmetric) การเปรียบเทียบผลการวิเคราะห์ในแต่ละควอนไทล์ซึ่งจะเป็นประโยชน์มากในการเข้าใจความสัมพันธ์ระหว่างตัวแปรตามและปัจจัยกำหนด ตามภาพที่ 4 กำหนดให้การแจกแจงของตัวแปรตาม Y ของแต่ละค่าของตัวแปรอิสระหรือตัวแปรกำหนด X ถ้าหากการแจกแจงของตัวรบกวนมีการแจกแจงแบบปกติ มีความเป็นอิสระไม่ขึ้นต่อกันและมีความแปรปรวนเท่ากัน จะทำให้เส้นถดถอยที่คำนวณได้ทั้ง 3 กรณีขนานกัน คือมีค่าความชันเท่ากัน จะแตกต่างกันตรงค่าคงที่หรือจุดตัด (intercept) ตัวแบบลักษณะนี้เรียกว่า "location shift model" หรือเรียกอีกชื่อว่า "homogeneous regression model" ในกรณีที่การแจกแจงที่มีความเบ้แต่ยังคงมีลักษณะการแจกแจงเหมือนกัน เส้นสมการถดถอยทั้งหมดที่ได้ยังคงขนานกัน แต่ระยะห่างของแต่ละเส้นควอนไทล์จะไม่เท่ากัน แต่ในกรณีที่ การแจกแจงของตัวรบกวนมีลักษณะการแจกแจงที่ไม่เหมือนกัน (non-identical conditional error distribution) เส้นสมการถดถอยที่ได้จะไม่ขนานกัน คือ มีความแตกต่างกันทั้งจุดตัดและความชัน ดังแสดงตามภาพที่ 5 ตัวแบบนี้เรียกว่า 'location and scales shift[4] model" สาเหตุที่ทำให้การแจกแจงที่ไม่เหมือนกัน คือ ค่าความแปรปรวนของตัวรบกวนมีค่าไม่เท่ากัน (heteroscedasticity) ตัวแบบนี้เรียกอีกชื่อว่า "heterogeneous regression model"

[3] ความเบ้ (skewness) เป็นกรณีที่เรียกว่า shape shift

[4] scale shift หมายถึงกรณีที่ค่าความแปรปรวนของตัวรบกวนมีค่าไม่คงที่ (heteroscedasticity)

รูปภาพที่ 5: สมการถดถอยที่ลากผ่านจุดเฉลี่ยและควอนไทล์ที่ 1 และ 3

กรณี "Heterogeneous regression model"



4. Quantile Regression Estimation[5]

การคำนวณหาค่าพารามิเตอร์ของสมการถดถอยภายใต้สมการถดถอยแบบ mean regression ใช้วิธีการคำนวณ Ordinary Least Squares ซึ่งมีหลักการคือ หาค่าพารามิเตอร์ที่ทำให้ผลรวมของค่าความคลาดเคลื่อนยกกำลังสองมีค่าต่ำสุด (minimize sum of squares of error terms) เขียนเป็นสมการดังนี้

[5] การคำนวณค่าพารามิเตอร์ของ Quantilie regression ทำได้ 2 แบบ คือ parametric and nonparametric

สมการถดถอย

$$y_i = a + bx_i + \varepsilon_i$$

$$min \sum_{n=1}^{n} (\mathrm{y_i} - (\mathrm{a} + bx_\mathrm{i}))^2$$

ค่าความคลาดเคลื่อน คือผลต่างระหว่างค่าสังเกตที่เกิดขึ้นจริงและค่าที่คำนวณได้จากสมการ (คือค่าเฉลี่ย (mean)) สำหรับหลักการของ QR ความแตกต่างระหว่างค่าสังเกตุและค่าที่คำนวณได้จากสมการวัดด้วย ผลรวมของน้ำหนักของผลรวมค่าสมบูรณ์ของความแตกต่าง (weighted sum of absolute value of vertical distances (differences)) โดยที่น้ำหนัก (weight) มีค่าเท่ากับ 1-p สำหรับจุดที่อยู่ใต้เส้นสมการถดถอย และ p เป็นจุดที่อยู่เหนือเส้นสมการถดถอย ซึ่งนำไปสู่การสร้างสมการหรือเส้นถดถอยที่เรียกว่า "conditional-quantile function" ดังแสดงในสมการที่ (1)

$$\sum_{i=1}^{n} d_p(y_i, \widehat{y_i}) = p \sum_{y_i \geq \beta_0^p + \beta_1^p x_i} \left[y_i - \beta_0^p - \beta_1^p x_i \right] + (1 - p) \sum_{y_i < \beta_0^p + \beta_1^p x_i} \left[y_i - \beta_0^p - \beta_1^p x_i \right] \ldots\ldots\ldots (1)$$

ทั้งนี้ d_p คือผลต่างระหว่างค่าคำนวณและค่าสังเกต ณ ระดับควอนไตล์ที่กำหนด ยกตัวอย่างค่า p อาจจะมีค่าเท่ากับ 0.10, 0.25 หรือ 0.50 เป็นต้น เช่น กรณีที่ p = 0.5 สมการที่ (1) เขียนได้ดังนี้

$$\sum_{i=1}^{n} d_p(y_i, \widehat{y_i}) = \sum \left[y_i - \beta_0^p - \beta_1^p x_i \right] \ldots\ldots\ldots (2)$$

สมการที่ (2) เรียกว่า Median regression เส้นสมการจะต้องลากผ่านจุด median ซึ่งจะมีจุดค่าสังเกตครึ่งหนึ่งอยู่เหนือเส้นและอีกครึ่งหนึ่งของค่าสังเกตอยู่ใต้เส้น ซึ่งมีหลายเส้น (หรือหลายๆ สมการ) เส้นที่ดีที่สุด คือ เส้นที่ทำให้ค่าที่คำนวณได้จากสมการที่ 2 มีค่าต่ำสุด

การคำนวณหาค่าพารามิเตอร์ของสมการถดถอยควอนไตล์ใช้ข้อมูลทั้งหมดของตัวอย่างที่ใช้ในการศึกษา ไม่ใช่แต่เฉพาะข้อมูลในส่วนของควอนไตล์ที่กำหนด สำหรับการคำนวณหาค่าพารามิเตอร์ β_0

และ β_1 ที่จะทำให้สมการควอนไทล์ที่ (1) มีค่าต่ำสุด นิยมใช้วิธีการโปรแกรมเชิงเส้นตรง (linear programming)

5. การประเมินผลการคำนวณของควอนไทล์รีเกรสชั่น

1. ประเมินผลการสอดคล้องกับข้อมูล (Goodness of Fit)

การวิเคราะห์แบบ MR จะพิจารณาจากค่า R^2 ซึ่งจะมีค่าอยู่ระหว่าง 0 กับ 1 ถ้า R^2 มีค่าเข้าใกล้ 1 มากเท่าใด แสดงว่า สมการ MR ที่คำนวณได้ปรับเข้ากับข้อมูลได้ดีมากขึ้น นั่นคือ สมการที่คำนวณได้สามารถอธิบายการเปลี่ยนแปลงของตัวแปรตามได้ดี การประเมินผลการสอดคล้องกับข้อมูลของสมการ QR จะเหมือนกับการประเมินผลสมการ MR คือผ่านค่า R^2 แต่สูตรในการคำนวณแตกต่างกัน จึงเรียกว่า "Pseudo R^2" ซึ่งมีค่าอยู่ระหว่าง 0 กับ 1 และให้ความหมายเหมือนกัน (อ่านรายละเอียดในส่วนนี้จาก Hao และ Naiman, 2007. หน้า 51-53)

2. ประเมินผลความมีนัยสำคัญของค่าสัมประสิทธิ์ของแต่ละสมการ QR

ประเมินผลโดยใช้หลักการเดียวกับการประเมินผลสมการ MR คือประเมินด้วยค่า t-statistics หรือค่า p= value

3. การทดสอบความมีนัยสำคัญของค่าสัมประสิทธิ์ของแต่ละสมการ QR

ด้วยกันหรือเปรียบเทียบกับ MR เป็นประเด็นที่สนใจ ทั้งนี้เพราะ ถ้าไม่มีความแตกต่างอย่างมีนัยสำคัญแสดงว่า การวิเคราะห์ที่แยกเป็นแต่ละ QR ก็ไม่มีประโยชน์ เพราะใช้ mean regression ตามปกติก็อธิบายได้ ดังนั้น การวิเคราะห์ QR จะมีประโยชน์ถ้ามีความแตกต่างกันของค่าพารามิเตอร์ของแต่ละสมการ QR (heterogenous in parameters) การทดสอบใช้ Wald test (อ่านเพิ่มเติมส่วนนี้ใน

เอกสารอ้างอิง

Davino, Cristina , Furno, Marilena and Vistocco, Domenico 2014. **Quantile Regression: Theory and Applications.** John Wiley and Sons

Koenher, Roger. 2005. **Quantile Regression.** Cambridge University Press

Hao, Lingxin and Naiman, Daniel. 2007. **Quantile Regression.** Sage Publications.