Robust Regression

1. Introduction

전통적인 통계적 방법들은 타당성을 인정받기 위해 여러 가지 모수적, 분포적 가정들을 필요로 한다. 그러나 현실에는 가정이 성립되지 않는 현상이나 자료들이 훨씬 많다. outlier(이상치)는 그러한 현상이나 자료들에 높은 확률로 존재하고 있다. outlier는 자료값 자체로서 존재할 수도 있고, 분포적 맥락에서 존재할 수도 있다. 즉 distributional outlier 역시 작지 않은 문제를 야기한다. robust method는 전통적인 통계적 추론 방법들이 갖는 한계를 극복하고자 태동하였고 그러한 방법들을 적용할 수 없을 경우 대안으로 시행하면 만족할 만한 결과를 제공한다.

특히 robust 회귀분석은 outlier를 갖는 자료, 특히 high leverage point를 갖는 선형 회귀 모형(linear regression model)에 적용할 수 있는 대표적인 방법으로 분석 결과가 outlier에 과도한 영향을 받지 않도록 해준다. 추정 결과가 outlier 쪽으로 치우치게 되면 추정값의 분산이 커져 통계적으로 매우 바람직하지 않다. 또한 전통적인 OLS(ordinary least squares) 추정 방법이 요구하는 여러 가지 가정이 성립하지 않을 경우에도 분석자가 원하는 결과를 얻지 못할 가능성이 크다. 우리는 여기서 robust 회귀에 대한 개괄적 소개를 통해 기본 개념과 주요 아이디어를 제공하는 데 초점을 맞출 예정이다. 우선 본론으로 들어가기 전 알아둬야 할 배경 지식들을 소개한다.

2. Preliminaries

2-1. Breakdown point

breakdown point란 쉽게 말해 추정량의 값을 임의로 변동시키지 않는 자료 비율을 의미한다. 평균을 예로 들어 설명하자면

$$\begin{aligned} \bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}\left[\sum_{i=1}^{n-1} x_i + x_n\right] \\ &= \frac{n-1}{n}\bar{x}_{n-1} + \frac{1}{n}x_n \end{aligned}$$

위의 세 번째 식에서 알 수 있듯이 x_n 하나만 임의의 큰 값을 가지게 되어도 표본평균 값 자체가 커지게 되므로 x_n은 $\bar{x}_n$에 영향을 미치는 값이다. 이것은 모든 $x_1, \ldots, x_n$에 대해서 마찬가지이므로 평균은 breakdown point of 0이다. 같은 맥락에서 중앙값은 breakdown point of 50%이다. breakdown point는 50%를 넘을 수 없고, higher breakdown point일수록 robust하다.

2-2. M-estimator

뒤에서 설명할 robust 회귀에서 가장 일반적으로 사용되는 추정 방법으로 1964년 Huber에 의해 처음 제안되었다. 우리가 통계적 추론에서 많이 사용하는 추정량들(평균, 중앙값, MLE 등)은 대부분 M-estimator의 special case이다. M-estimator는 robust함과 동시에 efficiency를 보장해 준다는 장점이 있고 방법 이름 자체가 'Maximum likelihood type'의 첫 글자를 딴 것에서 알 수 있듯 MLE 추정 방법의 일반적 형태이다. 간단히 정리하면 다음과 같다.

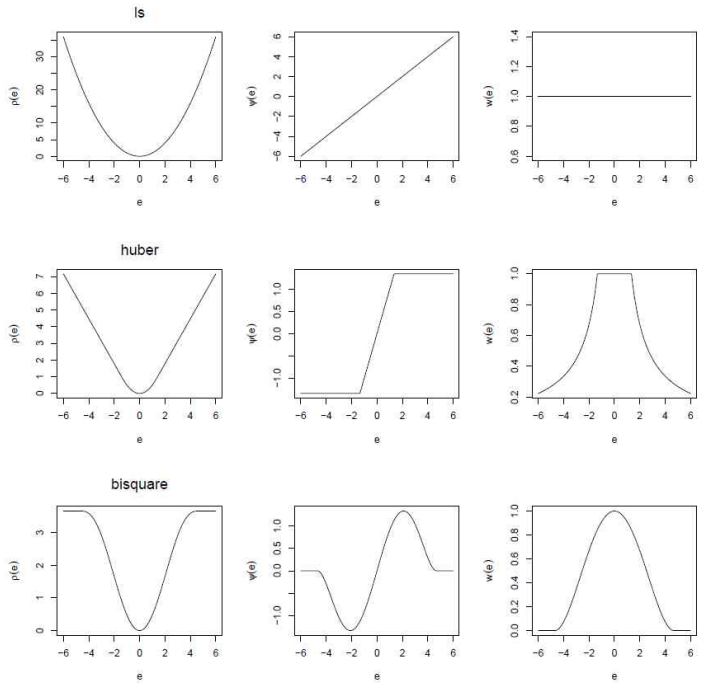
$$(M-estimator) = argmin_{\theta} \sum_{i=1}^{n} \rho(x_i, \theta)$$

여기서 $\rho(\bullet)$는 특정한 조건을 만족하는 임의의 함수를 의미한다.

Method	Objective Function	Weight Functions
Least Squares	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 \\ k\lvert e\rvert - \frac{1}{2}k^2 \end{cases}$	$w_H(e) = \begin{cases} 1 \\ \frac{k}{\lvert e\rvert} \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\{1-[1-(\frac{e}{k})^2]^3\} \\ \frac{k^2}{6} \end{cases}$	$w_B(e) = \begin{cases} [1-(\frac{e}{k})^2]^2 \\ 0 \end{cases}$

where $w(e) = \psi(e)/e$.

☞ 대표적인 objective function ρ인
least squares/Huber/bisquare objective function



☞ first column을 보면 ls보다는 Huber에서, Huber 보다는 bisquare일 때 outlier에 robust한 것을 확인할 수 있다. 또한 third column을 보면 ls의 경우는 모든 observation에 대하여 equal weight를 주는 반면, 나머지 두 경우에서는 0인 지점을 벗어나면 weight가 급격히 감소하는 것을 알 수 있다.

3. Robust Regression

3-1. Why Robust Regression?

robust 회귀 역시 전통적인 회귀 분석 방법의 한계를 극복하기 위해 등장하였다. 전통적으로 사용되어 오던 LSE(least squares estimation) 방법은 분포나 모수에 대한 여러 가정들이 필요하고 만약 그 가정들을 만족하지 못하면 잘못된 결과를 도출할 수 있다. LSE 방법은 outlier에 매우 민감하다. outlier가 단순히 정규분포의 꼬리부분에서 관측된 extreme value에 지나지 않는다면 아주 큰 문제는 아닐 수 있지만 그 outlier가 OLS(ordinary least squares) 하에서의 가정이 깨진 상황에서 관측된 것이라면 큰 문제가 발생하며 회귀분석 결과의 타당성을 저해할 수 있다. 그 외에 robust 회귀 방법은 이분산성(heteroscedasticity)이 존재하는 경우에도 분석에 장점을 갖는다.

가장 간단한 robust 회귀 방법으로는 LAD(least absolute deviation; L_1 regression) 방법이 있으나 앞서 설명했던 M-estimation 방법이 보다 포괄적이고도 일반적인 방법으로 널리 쓰이고 있다. M-estimation 방법을 통한 모수 베타(β)의 추정 과정은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i$$

위와 같은 선형 모형이 존재할 때

$$\begin{aligned} \hat{\beta} &= \underset{\beta}{argmin} \sum_{i=1}^{n} \rho(e_i) \\ &= \underset{\beta}{argmin} \sum_{i=1}^{n} \rho(y_i - x_i^T \hat{\boldsymbol{\beta}}) \end{aligned}$$

여기서 ρ는 특정한 조건을 만족하는 임의의 함수를 의미한다. 위와 같은 방법으로 추정한 모수 베타는 outlier에 robust한 성질을 갖는다. 한편 $'\rho(e) = e^2'$의 2차 함수를 사용할 경우에는 OLS 방법과 동일하다.

3-2. Parametric approach to Robust Regression

robust 회귀에서 위와 같이 M-estimation 방법을 적용하는 것 외 또 하나의 대안은 error term의 정규분포 가정을 다른 'heavy-tailed distribution'으로 대체하는 것이다. 현실에서의 다양한 경우에서 자유도 4~6 사이의 t 분포는 실제로 좋은 선택지를 제공하고 있다. 특히 Bayesian robust regression에서는 이러한 분포들에 강하게 의존하는 경향이 있다. 또 하나의 방법으로는 잔차(residual)가 혼합정규분포(Gaussian mixture distribution)를 따른다는 가정을 두는 것이다. 즉,

$$e_i \sim (1-\varepsilon)N(0, \sigma^2) + \varepsilon N(0, c\sigma^2).$$

여기서 $c > 1, \epsilon < 0.1$이다. 위와 같이 contaminated normal distribution을 가정하여 분석을 진행하는 경우도 흔히 있다. parametric approach는 통계적 추론 시 likelihood function을 이용한 규격화된 접근을 가능하게 하여 적절한 시뮬레이션 모델을 세우기에 용이하다는 장점이 있지만, 결국 분포에 대한 가정이 필요하다. 또한 skewed residual distribution의 경우에는 만족할 만한 설명을 제공해주지 못한다.

4. Kernel of Robust Regression : Quantile Regression

quantile regression이란, 용어 그대로 반응변수 Y의 '분위수(quantile)'에 관심을 갖는 회귀 분석 방법이다. 선형 회귀 모형에서 특히 많이 사용되며 asymmetric conditional distribution을 가질 때 전통적 방법에서의 mean function보다 quantile function은 더 설명력을 갖는다. quantile regression의 장점과 단점은 다음과 같다.

※ Pros and cons of quantile regression

◇ Advantage
① 자료를 보다 완전한 시각으로 볼 수 있다.
② outlier에 robust하다.
③ error term의 normality가 깨져있는 경우 OLS에 비해 더 효율적인 추정 방법이다.
④ outcome의 original measurement scale대로 해석이 가능하다.

◇ Disadvantage
① quantile set에 대한 회귀적합이 필요하므로 computationally intensive하다.
② solution in closed form이 존재하지 않는다.
③ non-continuous case에 적용하기 힘들다.

quantile을 다음과 같이 정의할 때,

$$Q_Y(\tau) = F_y^{-1}(\tau) = Inf\{y : F_Y(y) \geq \tau\}, \tau \in (0,1)$$

OLS 방법에서의 conditional mean function과 quantile regression 방법에서의 conditional quantile function을 비교하면 다음과 같다.

1. Conditional mean of Y given $X = x$

$$E(Y|X = x) = \underset{\alpha}{argmin} E[(Y - \alpha)^2 | X = x]$$

2. Conditional quantile of Y given $X = x$

$$q_\tau(x) = Q_{Y|X}(\tau|X = x) = \underset{\alpha}{argmin} E[\rho_\tau(Y - \alpha)|X = x]$$

여기서 $\rho_\tau(y) = y(\tau - I(y < 0))$, I는 indicator function이다. τ가 0.5일 경우 median regression이라고도 불리는 L_1 regression 형태를 가지며 이때는 $x_i^T\beta$의 중앙값을 찾는 문제로 치환된다. 정리하면,

$$\boldsymbol{\beta_\tau} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{argmin} E(\rho_\tau(Y - \boldsymbol{X\beta}))$$

위의 과정을 통해 추정된 모수로 다음의 conditional quantile function을 얻는 것이다.

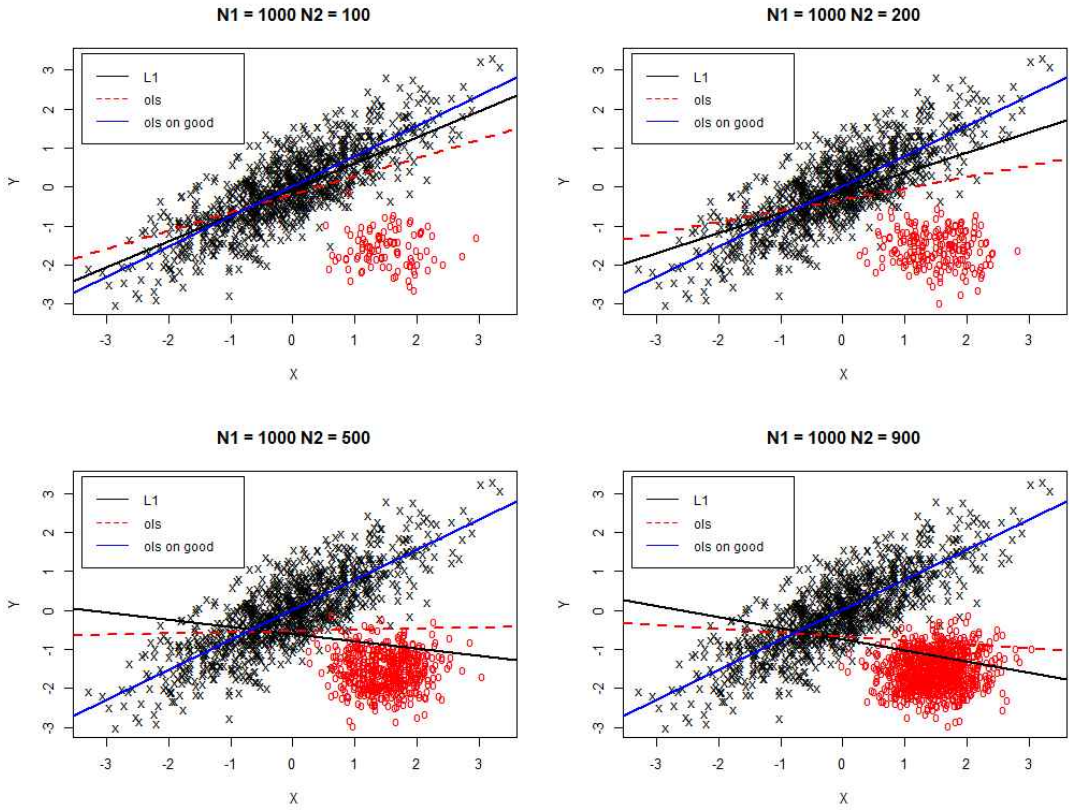
$$Q_{Y|X}(\tau) = \boldsymbol{X\beta_\tau}$$

5. Simulation using Real Data

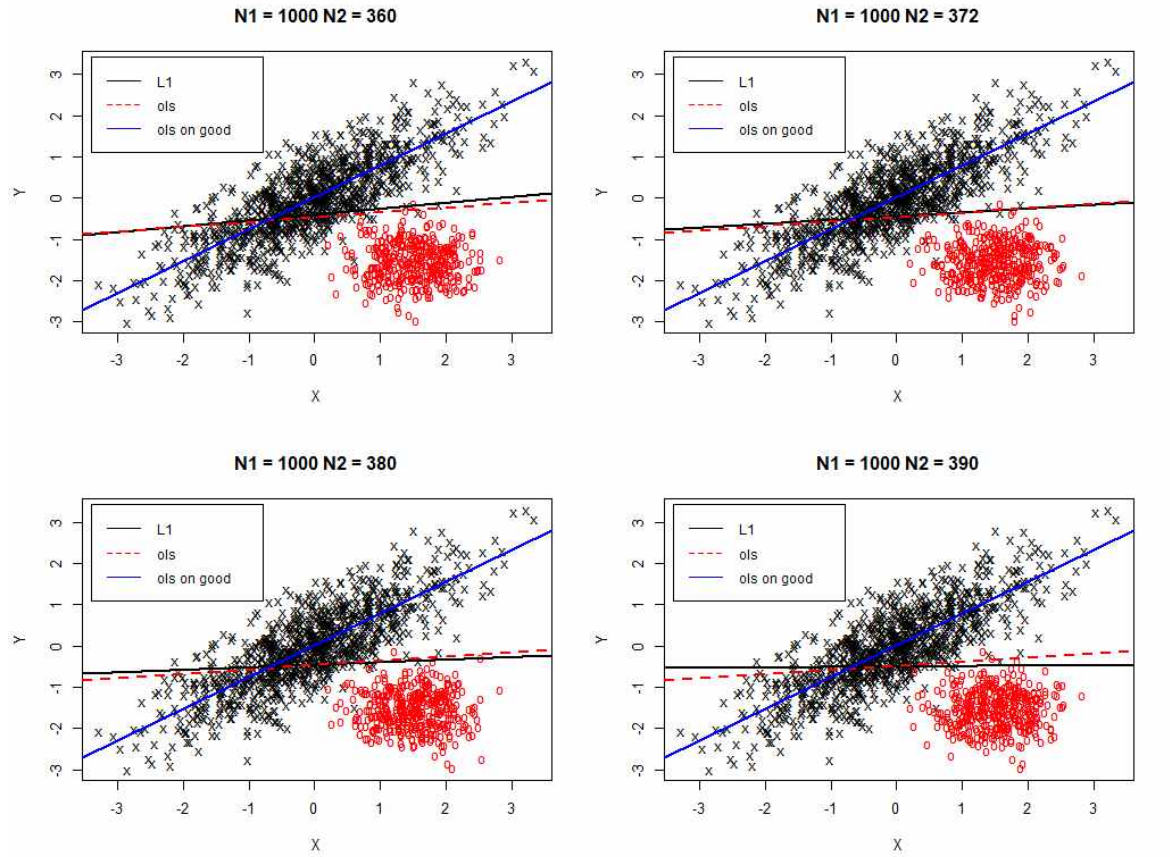
5-1. L_1 Regression

$$\tilde{\beta} = argmin \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i^T \beta|$$
$$= argmin \frac{1}{n} \sum_{i=1}^{n} \rho_{0.5}(y_i - x_i^T \beta)$$

L_1 regression(=median regression)과 OLS regression의 performance 비교를 위해 다음과 같은 경우를 생각하였다.



위의 경우는 performance의 비교를 위해 가장 간단하고도 극단적인 data를 상정하여 분석해본 것이다. n_1개의 'good' data와 n_2개의 'bad' data가 공존하는 경우 bad data의 크기가 커질수록 각각 L_1 회귀에서의 fitted line과 OLS 회귀에서의 fitted line이 어떤 trend를 보이는지 살펴보았다. 빨간색 점선이 OLS case에서의 fitted line이고 검은색 실선이 L_1 case에서의 fitted line이다. 플롯에서 확인할 수 있는 바와 같이 bad data의 크기가 커질수록 두 fitted line은 공통적으로 bad data 쪽으로 치우친 채 적합이 된다. 그러나 한 가지 주목할 점은 bad data의 크기가 극단적으로 큰 경우에는 오히려 L_1 fitted line이 OLS fitted line보다 더 bad data 쪽으로 치우치는 문제가 발생한다. 다음의 플롯을 살펴보자.



위의 플롯들은 L_1 fitted line이 OLS fitted line보다 bad data에 robust 하도록 하기 위한 threshold n_2 size에 대한 것이다. n_2 size가 360보다 더 큰 어떤 지점을 넘어서면 오히려 L_1 regression 방법이 더 bad data에 sensitive하게 되는 것이다. 물론 위의 경우는 분석과 해석의 편의를 위해 설정한 극단적인 경우이다. data의 분포가 위에서 주어진 scatterplot과 같은 경우는 다른 변수를 고려하거나 mixture distribution을 상정하여 생각할 수도 있다. outlier(여기서는 bad data)의 비중이 유의미하게 높지 않다면 L_1 regression은 합리적인 대안이다.

5-2. Quantile Regression

quantile regression 방법을 사용하는 주요 목적은 크게 두 가지로 나눌 수 있다.

● 이분산성(heteroscedasticity) 식별

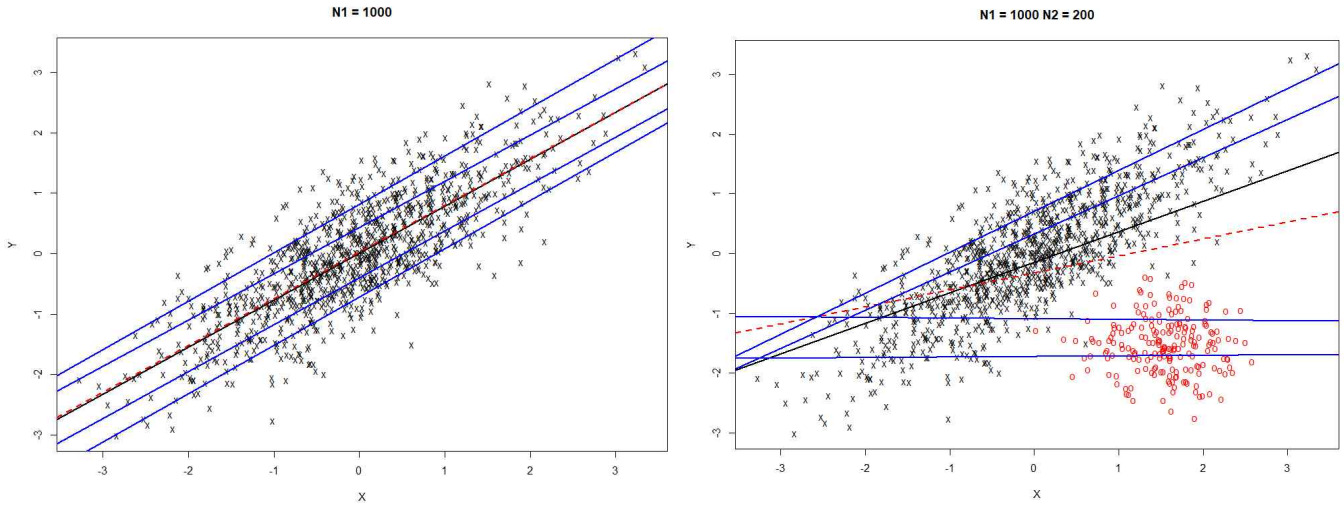
● 예측구간(prediction interval) 추정 : 예측구간은 두 종류를 추정할 수 있다.

① $[q_{\alpha/2}(x), q_{1-\alpha/2}(x)]$: raw data의 scatterplot에서 그린 quantile regression line을 이용해 각 x변수값에 대응하는 y변수값의 quantile로 구간을 추정

② linear quantile regression의 회귀계수(β_0, β_1)에 대한 구간 추정

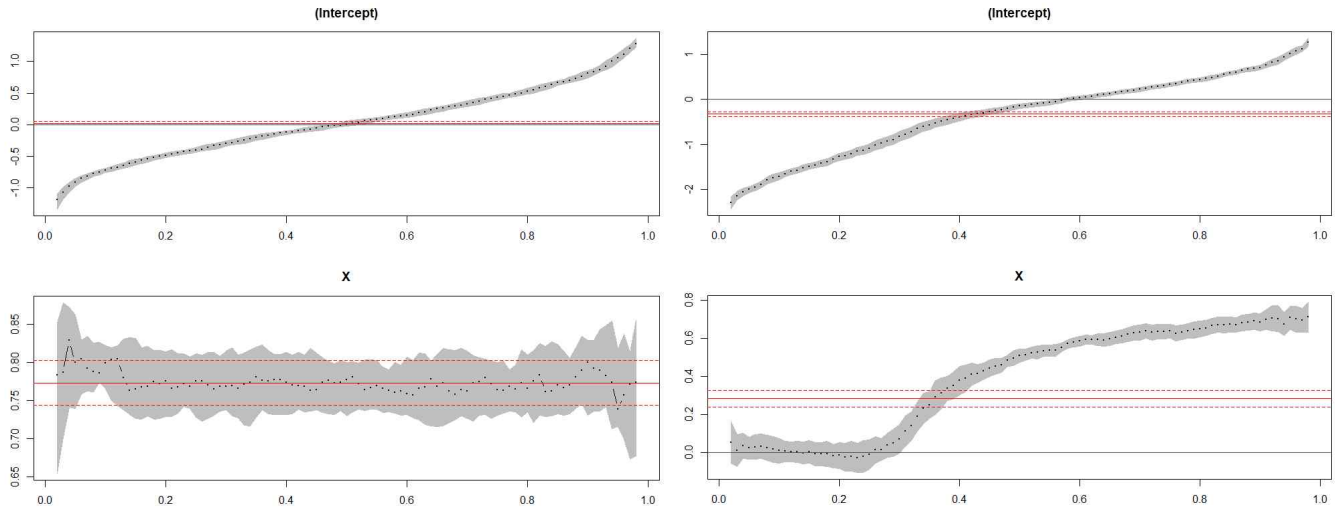
각각의 경우를 플롯을 통해 확인해보기로 한다. (아래 플롯들의 좌우는 각각 다른 경우이다.)

(이어짐)

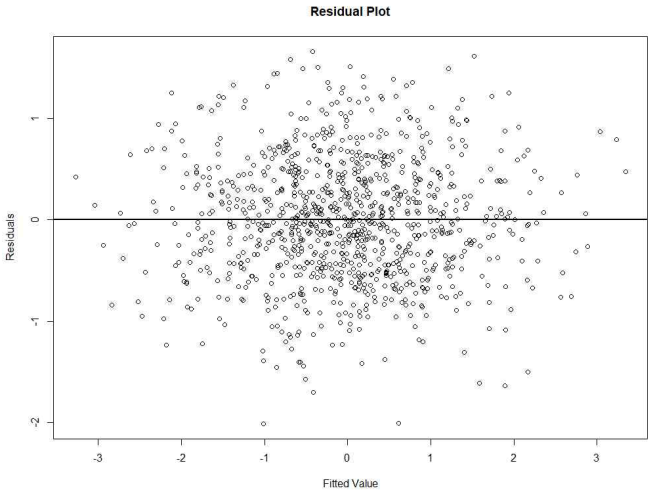


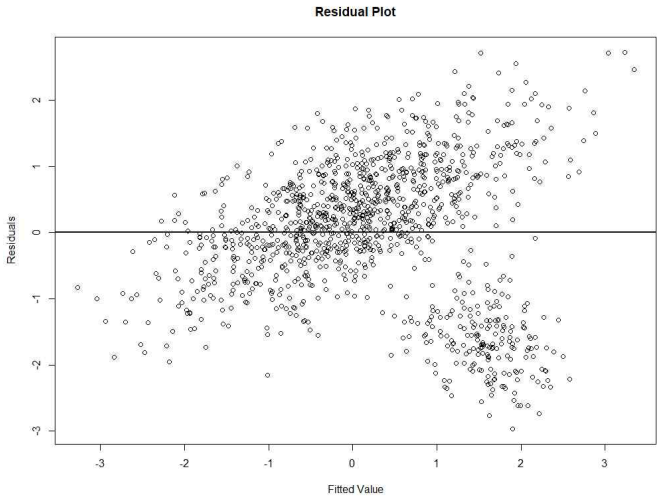
이분산성(heteroscedasticity)은 quantile function을 그려보면 상대적으로 쉽게 식별해낼 수 있는 성질이다. 결론적으로 왼쪽 플롯은 이분산성이 거의 존재하지 않는 경우, 오른쪽은 이분산성이 존재하는 경우이다. 두 플롯에는 각각 10%, 25%, 50%, 75%, 90% quantile regression line들이 그려져 있다(검은 선은 median regression line). 왼쪽 플롯을 보면 5개의 quantile regression line이 겹치지 않고 거의 평행을 유지하고 있다. 그러나 오른쪽 플롯의 quantile regression line들은 평행하지 않고 기울기가 다르며 오른쪽으로 갈수록 넓게 퍼지는 모양을 갖고 있다.

또한 quantile regression line 사이의 영역은 각 x변수값에 대응하는 y변수값의 quantile로 추정된 구간이라고 볼 수 있다. 예를 들어 10% line과 90% line 사이의 영역은 각 x변수값에 대한 80% prediction interval이 되는 것이다.



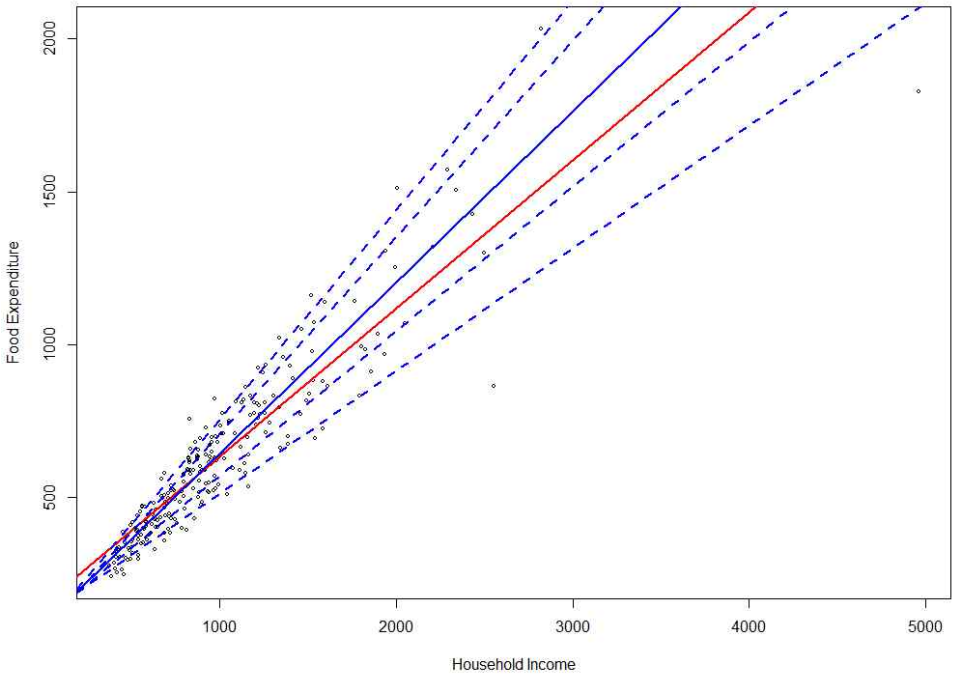
왼쪽과 오른쪽의 두 플롯은 각각 바로 위에 있는 scatterplot들에 대응되며, 검은 점선과 회색 영역은 linear quantile regression model에서의 intercept coefficient, slope coefficient에 대한 각 τ에서의 estimate 값들과 그때의 95% prediction band를 그린 것이다. 빨간색 실선과 점선은 각각 OLS estimate 값과 95% confidence interval을 나타낸다. 왼쪽 아래 플롯을 보면 quantile regression에서의 estimate 값 & prediction band와 OLS estimate 값 & confidence interval이 거의 일치함을 알 수 있다. 이를 통해 OLS의 accuracy를 상대적으로 체크해볼 수 있음과 동시에 prediction interval을 통해서도 '이분산성 or 등분산성' 여부를 확인할 수 있다.



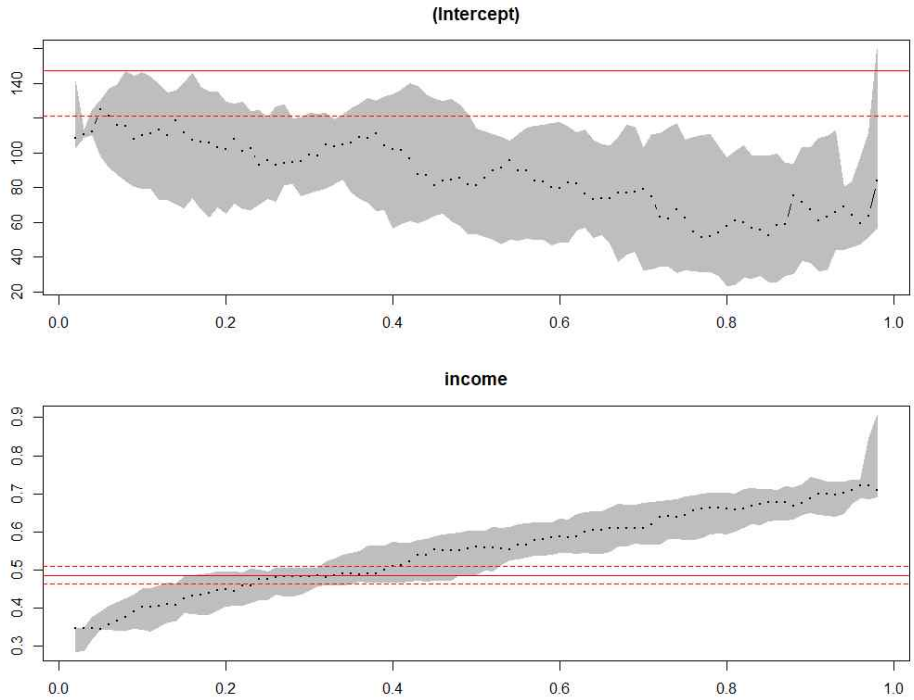


위의 플롯은 결과 확인을 위해 각각의 경우에 raw data로 linear regression을 적합한 후 그린 residual plot이다. 왼쪽 플롯을 보면 자료점의 분포에 특별한 패턴이 보이지 않는다. 그러나 오른쪽 플롯을 보면 residual plot의 자료점들이 두 개의 group으로 분류가 되는 패턴을 보인다. 그러므로 오른쪽 경우의 raw data에는 이분산성이 존재함을 알 수 있다.

다음 예제는 quantile regression을 설명할 때 다루는 대표적인 data로 19세기 벨기에 가정의 소득과 식비 지출 간의 관계에 대한 data(built-in data in R)를 이용한 것이다.



위의 플롯을 보면 앞에서 봤던 예제와 같이 quantile regression line(10%, 25%, 50%, 75%, 90%, 이 중 median regression line은 파란색 실선)이 평행하지 않고 오른쪽으로 갈수록 넓게 퍼지는 모양이다. data에 이분산성이 존재함을 확인할 수 있다. linear quantile regression에서의 intercept coefficient estimate, 'income' 변수에 대한 slope coefficient estimate와 prediction band는 다음과 같다.



다음은 각 τ에 대하여 변수 'income'의 coefficient estimate와 prediction interval을 나타낸 것이다.

Quantiles	estimate & P.I.
0.05	0.343 (0.343, 0.390)
0.25	0.474 (0.420, 0.494)
0.50	0.560 (0.487, 0.602)
0.75	0.644 (0.580, 0.690)
0.95	0.709 (0.674, 0.734)