BAYESIAN QUANTILE REGRESSION FOR LONGITUDINAL STUDIES WITH NONIGNORABLE MISSING DATA

PAPER REVIEW

Original by
Ying Yuan and Guosheng Yin

Published In
Biometrics

Review by
202071296 권나연, 2020712421 임현우

December 2021

Contents

1 서론

경시적자료에 대한 연구는 종종 missing data가 무시할 수 없거나 정보를 가진 경우 model parameter의 편향된 추정으로 인해 어려움을 겪는다. 이를 처리하기 위해서 다양한 모델들이 많이 나왔으나, 이러한 모델들은 대부분 평균회귀분석(mean regression)을 기반한 모델이다(Wu and Carroll, 1988; Wu and Bailey, 1989; Little, 1993; Diggle and Kenward, 1994; among others). 그에 반해 QR(Quantile Regression)모델은 non-ignorable missing data를 처리하는데에 많은 제한이 존재한다. 이 논문은 common random effects를 QR모델과 missing data model에 공유함으로써 non-ignorable missing data를 처리하는 shared-parameter QR model을 제시한다.

먼저 본 논문에서 다룬 데이터는 Pediatric AIDS Data이며, 지도부딘의 lower dosage 와 higher dosage 의 효과를 비교하기 위해 고안되었다. 총 424개의 subject가 있으며 그중 216개를 low-dose , 208개를 high-dose 그룹으로 랜덤하게 나눈 후 CD4 세포수를 매 12주마다 200주 까지 수집하였다. HIV환자들의 경우 CD4 세포수가 적을수록 더 위험하므로, CD4 세포수가 더 느리게 감소한다는 것은 그 Treatment가 더 좋은 효과임을 나타낸다. 그러나 Pediatric AIDS Data는 간헐적으로 missing data와 dropout이 많다는 문제점이 있다. Low-dose 그룹에서는 52% , 그리고 high-dose그룹에서는 45% 만이 3년간의 수집을 완료하였으며 심지어 각 subject 50%이상이 하나이상의 missing data가 존재했다. 에이즈 연구에서 이러한 missing data는 CD4 세포수와 관련있기 때문에 무시할 수 없는 데이터이다(Wu and Carroll, 1988; De Gruttola and Tu, 1994; and Hogan and Laird, 1997).

다음은 Pediatric AIDS Data에 median regression model을 각 subject에 대해 적용한 결과이다.

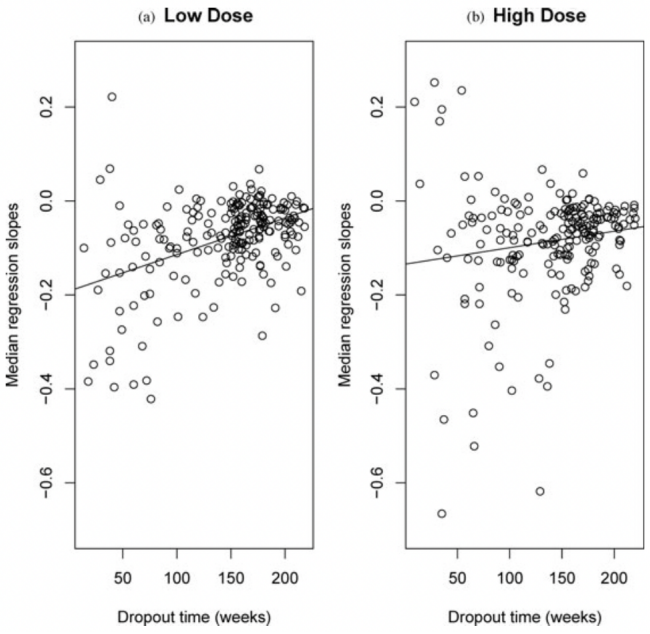

Figure 1: Individual median regression slopes for the square root of the CD4 cell count versus dropout times.

Figure 1에서 볼 수 있듯이, 기울기가 더 낮을수록 더 빨리 dropout을 하는것을 확인 할 수 있다. 이때, 기울기는 CD4 세포수에 대한 기울기이다. 이는 missing data가 informative하다는 것을 의미하므로, missing data를 무시할 수 없다는 것을 의미한다.

이러한 non-ignorable missing data를 처리하기 위해서, 일반적인 QR 모델의 check function에 penalize

를 줌으로써 각 level의 QR parameter를 모집단 값으로 shrunk하는 shared-parameter QR model의 사용을 제안한다. 이때 likelihood framework에서 penalized check function을 random effects model로 전환하는 과정을 거친다. 또한 shared latent subject-specific random effects를 통해 outcome process와 missing data process가 관련있다고 가정한다. 그리고 Bayesian paradigm에서 MCMC(Markov chain Monte Carlo)를 이용하여 추정과 추론을 진행한다.

2 본론

2.1 Methods

2.1.1 Quantile Regression

y_i를 outcome, x_i를 그에 해당하는 covariate vector라고 표기한다.이때 τ번째 QR model은 다음과 같다.

$$Q_{y_i}(\tau|x_i) = x_i^T\beta \text{ , where } 0 < \tau < 1.$$

Regression parameter β를 다음과 같은 check function을 minimizing함으로써 추정한다. 이때 check function은 ALD와 밀접한 관련이 있다.

$$\sum_{i=1}^n \rho_\tau(y_i - x_i^T\beta), \text{ where } \rho_\tau(u) = u\{\tau - I(u < 0)\}.$$

또한 scale parameter가 1인 ALD의 확률밀도 함수는 다음과 같다.

$$f(y|\mu, \tau) = \tau(1-\tau)exp\{-\rho_\tau(y-\mu)\}.$$

이때 τ는 quantile level, μ는 location parameter이다. 위 식을 통해 ALD 확률밀도함수에 check function이 포함되어 있음을 확인할 수 있다. y_i가 $\mu = x_i^T\beta$인 ALD를 따른다고 가정한다면 check function을 minimizing하는 것과 y의 likelihood를 maximizing하는 것이 동일하다는 것을 확인할 수 있다. 이러한 관계를 이용하여 likelihood framework를 reformulate한다.

2.1.2 Modeling Longitudinal Data

n개의 subject와 미리 지정된 J개의 시점에서 반복적으로 측정하는 longitudinal study를 고려하고자 한다. y_{ij}는 i번째 subject의 j시점에서의 outcome을 나타낸다. τ번째 QR에서, 다음과 같은 l_2-penalized check function을 제안한다.

$$\sum_{i=1}^n \sum_{j \in J_{i,obs}} \rho_\tau(y_{ij} - x_{ij}^T\beta - z_{ij}^T b_i) + \frac{1}{2}\sum_{i=1}^n b_i^T \Lambda^{-1} b_i.$$

이때 x_{ij}와 z_{ij}는 covariates이고, b_i는 subject-specific effects, Λ는 symmetric nonsingular matrix이다. 이러한 penalty term을 사용함으로써 b_i를 0 쪽으로 shirink할 수 있으며 shirink의 양은 Λ를 tunning함으로써 조절한다. l_2-Penalized check function은 다음과 같은 모델을 따르는 likelihood framework에 casting 할 수 있다.

$$y_{ij}|b_i \sim ALD(\tau, x_{ij}^T\beta + z_{ij}^T b_i),$$
$$b_i \sim N(0, \Lambda)$$

위와 같은 분포가정을 통해 l_2-penalized check function을 최소화 하는 것과 likelihood를 최대화 하는 것을 동일하게 만들어 준다.

2.1.3 Modeling Non-Ignorable Missing Data

Non-ignorable missing data를 설명하기 위해, missing data와 dropout process를 모델링하고, 이를 outcome process에 연결하려 한다. Missing data status s_ij를 다음과 같이 정의한다.

s_{ij} = $\mathcal{O}$ if measurement j of subject i is observed.
s_{ij} = $\mathcal{I}$ if measurement j of subject is intermittent missing.
s_{ij} = $\mathcal{D}$ if subject i drops out at measurement j.

이때, 모든 subject가 J=1시점에서 측정되어있으며 dropout된 이후에는 측정하지 않는다고 가정한다. 이제 same random effects인 b_i를 공유한다고 가정함으로써 missing data process와 outcome process를 연결한다. 특히 missing data process를 다음과 같은 transition probability를 사용하여 모델링 한다.

$$\begin{aligned}
\pi_{ij}^{\mathcal{O}} &= Pr(s_{ij} = \mathcal{O} | s_{i(j-1)} \neq \mathcal{D}, b_i) = \frac{1}{1 + \sum_{k \in (\mathcal{I}, \mathcal{D})} exp(w_{ij}^T \alpha^{(k)} + b_i^T \gamma^{(K)})} \\
\pi_{ij}^{\mathcal{I}} &= Pr(s_{ij} = \mathcal{I} | s_{i(j-1)} \neq \mathcal{D}, b_i) = \frac{exp(w_{ij}^T \alpha^{(\mathcal{I})} + b_i^T \gamma^{(\mathcal{I})})}{1 + \sum_{k \in (\mathcal{I}, \mathcal{D})} exp(w_{ij}^T \alpha^{(k)} + b_i^T \gamma^{(K)})} \\
\pi_{ij}^{\mathcal{D}} &= Pr(s_{ij} = \mathcal{D} | s_{i(j-1)} = \mathcal{O}, b_i) = \frac{exp(w_{ij}^T \alpha^{(\mathcal{D})} + b_i^T \gamma^{(\mathcal{D})})}{1 + \sum_{k \in (\mathcal{I}, \mathcal{D})} exp(w_{ij}^T \alpha^{(k)} + b_i^T \gamma^{(K)})}
\end{aligned} \tag{1}$$

이때 w_{ij}는 covariates, $\alpha^{(k)}$는 그에 해당하는 regression parameter, $\gamma^{(k)}$ 는 b_i와 missing data process 간의 관계를 제어해주는 parameter를 나타낸다. Subject i의 missing data process의 log conditional likelihood는 다음과 같이 표현된다.

$$\sum_{j=2}^{J} log f(s_{ij} | b_i) = \sum_{j=2}^{J} \{ I(s_{ij} = \mathcal{O}) log \pi_{ij}^{(\mathcal{O})} + I(s_{ij} = \mathcal{I}) log \pi_{ij}^{(\mathcal{I})} + I(s_{ij} = \mathcal{D}) log \pi_{ij}^{(\mathcal{D})} \}.$$

이 경우 subject i가 dropout된 이후에 s_{ij}는 정의되지 않는다. 위와 같이 b_i를 outcome process와 공유함으로써 missing data model은 non-ignorable missing data를 설명할 수 있게 된다.

2.1.4 Posterior Estimation

y,s를 관측데이터라고 하면 관측데이터에 대한 likelihood는 다음과 같이 표현할 수 있으며, 앞서 정의한 모델들을 이용하여 표현할 수 있다.

$$L(y, s | \beta, \Lambda, \alpha^{(K)}, \gamma^{(k)}) = \prod_{i=1}^{n} \int \prod_{j \in J_{i,obs}} f(y_{ij} | b_i) \prod_{j=2}^{J} f(s_{ij} | b_i) f(b_i) db_i, \; k = \mathcal{I}, \mathcal{D}$$

베이지안 관점에서, $P(\beta, \Lambda, \alpha^{(k)}, \gamma^{(k)})$는 prior distribution이며 joint posterior distribution은 베이즈 법칙에 의해 다음과 같이 표현된다.

$$P(\beta, \Lambda, \alpha^{(K)}, \gamma^{(K)} | y, s) \propto L(y, s | \beta, \Lambda, \alpha^{(K)}, \gamma^{(K)}) P(\beta, \Lambda, \alpha^{(K)}, \gamma^{(K)}).$$

또한 $\beta, \alpha^{(k)}, \gamma^{(k)}$에는 non-informative-prior를 , Λ^{-1}에는 Wishart 사전분포를 가정한다.

$$\begin{aligned}
\beta, \alpha^{(K)}, \gamma^{(K)} &\propto 1, \; k = \mathcal{I}, \mathcal{D} \\
\Lambda^{-1} &\sim WI(q, cI)
\end{aligned}$$

WI(q,cI)는 자유도 q와 scale matrix cI를 가지는 Wishart 분포임을 의미하며 이때 c는 작은 상수, I는 단위 행렬이다.
얻어낸 joint posterior distribution을 통해 각 parameter의 full conditional distribution을 계산하고, Gibbs sampler를 이용하여 posterior distribution을 얻는다. 하지만 Λ를 제외하고는 parameter의 full conditional distribution은 closed form을 가지지 않는다. 이 경우에는 adaptive Metropolis Hasting sampling 알고리즘을 이용하여 parameter의 사후분포를 얻는다. 이후에 얻어낸 사후분포와 Sample을 이용하여 추정과 추론을 진행한다.

2.2 Application

Application에서는 QR을 Pediatric AIDS Data에 적용하여 데이터 분석을 진행하였다. 여기서 t_{ij}는 i번째 subject의 j번째 측정시간, y_{ij}는 t_{ij}에서 측정된 CD4 세포수의 제곱근, 그리고 x_i는 binary treatment indicator이다. 만약 $x_i = 0$라면 약을 고용량으로 투여함을 의미한다. τ번째 regression quantile에서 다음의 회귀모형을 고려할 수 있다.

$$\begin{aligned} &Q_{y_{ij}}(\tau \mid x_i, t_{ij}, b_{0j}, b_{1i}) \\ &= \beta_0 + \beta_1 x_i + \beta_2 t_{ij} + \beta_3 x_i t_{ij} + b_{0j} + b_{1i} t_{ij}, \end{aligned} \tag{2}$$

where the β's characterized the population-level trajectory, $b_{0i} \sim N(0, \lambda_0)$ and $b_{1i} \sim N(0, \lambda_1)$

그래서 missing data model은 다음과 같이 나타난다.

$$\begin{aligned} \pi_{ij}^{(o)} &= \frac{1}{1 + \sum_{k \in (I,D)} exp(\gamma_0^{(k)} + \gamma_1^{(k)} b_{0i} + \gamma_2^{(k)} b_{1i} + \gamma_3^{(k)} x_i + \gamma_4^{(k)} x_i b_{0i} + \gamma_5^{(k)} x_i b_{1i})} \\ \pi_{ij}^{(I)} &= \frac{exp(\gamma_0^{(I)} + \gamma_1^{(I)} b_{0i} + \gamma_2^{(I)} b_{1i} + \gamma_3^{(I)} x_i + \gamma_4^{(I)} x_i b_{0i} + \gamma_5^{(I)} x_i b_{1i})}{1 + \sum_{k \in (I,D)} exp(\gamma_0^{(k)} + \gamma_1^{(k)} b_{0i} + \gamma_2^{(k)} b_{1i} + \gamma_3^{(k)} x_i + \gamma_4^{(k)} x_i b_{0i} + \gamma_5^{(k)} x_i b_{1i})} \\ \pi_{ij}^{(D)} &= \frac{exp(\gamma_0^{(D)} + \gamma_1^{(D)} b_{0i} + \gamma_2^{(D)} b_{1i} + \gamma_3^{(D)} x_i + \gamma_4^{(D)} x_i b_{0i} + \gamma_5^{(D)} x_i b_{1i})}{1 + \sum_{k \in (I,D)} exp(\gamma_0^{(k)} + \gamma_1^{(k)} b_{0i} + \gamma_2^{(k)} b_{1i} + \gamma_3^{(k)} x_i + \gamma_4^{(k)} x_i b_{0i} + \gamma_5^{(k)} x_i b_{1i})} \end{aligned} \tag{3}$$

그리고 Markov chain의 수렴을 확인하기 위해 the Gelman–Rubin convergence statistic (shrinkage factor)을 계산하였다. 초기 1,000번을 burn-in 후의 이 shrinkage factor의 값은 1에 가까운 값이 나왔기 때문에 이 체인은 수렴한다고 할 수 있다.

		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
QR model		Est.	95% CI	Est.	95% CI	Est.	95% CI
Random effects	β_0	21.40	(20.40, 22.41)	23.62	(22.56, 24.68)	25.97	(24.86, 27.08)
	β_1	0.14	(−1.84, 2.11)	0.19	(−1.94, 2.31)	0.24	(−2.01, 2.47)
	β_2	−3.19	(−3.54, −2.86)	−3.54	(−3.89, −3.20)	−3.82	(−4.19, −3.46)
	β_3	0.78	(0.11, 1.45)	0.82	(0.13, 1.51)	0.83	(0.127, 1.55)
	λ_0	104.49	(90.94, 120.45)	117.98	(102.61, 135.76)	129.61	(112.56, 149.17)
	λ_1	9.49	(7.83, 11.38)	10.67	(8.86, 12.75)	11.09	(9.14, 13.36)
Shared parameter	β_0	21.09	(20.07, 22.08)	23.33	(22.24, 24.39)	25.58	(24.45, 26.70)
	β_1	0.17	(−1.86, 2.14)	0.21	(−1.92, 2.28)	0.27	(−1.98, 2.47)
	β_2	−3.40	(−3.76, −3.06)	−3.75	(−4.11, −3.40)	−4.09	(−4.47, −3.72)
	β_3	0.77	(0.06, 1.44)	0.81	(0.08, 1.52)	0.84	(0.11, 1.57)
	λ_0	105.87	(92.09, 121.59)	119.77	(104.40, 137.65)	132.02	(114.93, 151.51)
	λ_1	9.55	(7.92, 11.41)	10.95	(9.10, 13.12)	11.69	(9.68, 14.04)

Figure 2: Estimates and 95% CIs of the model parameters for the pediatric AIDS data

위의 Figure 2는 각각 regression quantile인 τ가 0.25, 0.5, 그리고 0.75일 때 shared-parameter QR model과 random-effects QR model 하에서의 model parameter들의 추정값들을 보여준다. τ의 값이 낮을 수록 CD4 세포수가 적으며, 이는 중증환자임을 뜻한다. β_1의 추정값은 세 분위수에서 모두 0에 가까운 값이었으며, 이는 두 treatment 간에 baseline CD4 세포수의 밸런스가 잘 맞는다고 해석할 수 있다. 이는 randomization에 의하여 나타난 결과이다. Shared-parameter QR model 하에서의 β_2의 추정값은 random-effect QR 하에서의 β_2의 추정값보다 더 작았다. 이는 shared-parameter QR에서 초기 dropout이 더 낮은 기울기와 연관되어 있다는 사실을 고려하기 때문이다. 예를 들어, 중앙값 β_2의 추정값을 살펴보면 약을 고용량으로 투여했을 때의 시간에 대응하는 기울기의 추정값은 random-effects QR model의 경우 -3.54, shared-parameter QR model의 경우 -3.75였다. Shared-parameter QR model의 경우 고용량과 저용량으로 약을 투여한 경우 모두 더 위독한 환자에게 더 효과적이었다. 제1 사분위수의 경우 β_2의 추정값이 -3.40이며, 제3 사분위수의 경우 β_2 추정값이 -4.09다. 이는 위독한 환자들의 CD4 세포수가 더 느리게 줄어들었다는 것을 의미한다. β_3의 추정값은 두 treatment 간의 시간의 기울기 차이를 의미한다. 두 모델 모두 유사한 결과가 나타났지만, 저용량으로 약을 투여한 경우의 시간의 기울기가 고용량으로 약을 투여한 경우보다 훨씬 높았다. $\beta_3 > 0$이었는데, 이는 저용량으로 약을 투여하는 것이 고용량으로 투여하는 경우보다 효과적이라는 것을 의미한다. 하지만, 저용량으로 약을 투여했을 때의 우수성은 낮은 분위수의 경우 덜했다. β_3의 추정값이 제1 분위수의 경우 0.77이었지만, 제3 사분위수는 0.84로 제3 사분위수가 조금 더 높은 값인 것을 통해 알 수 있다.

3 결론

본 논문은 무시할 수 없는 간헐적인 missing data와 dropout이 존재하는 경시적 데이터를 위한 QR을 제시하였다. 또한 ℓ_2 regularization을 사용하여 subject-specific regression line를 population line으로 shrink하였다. 이는 within-subject correlation을 설명한다. 그리고 missing data process는 공통된 random effects 공유를 통해서 longitudinal outcome process와 관련이 있다고 가정하였다. QR check function과 ALD 간의 관계를 활용함으로써, QR problem을 보통의 likelihood framework로 보였다. 그리고 모델의 모수와 분산의 posterior 추정값을 제공하는 Bayesian MCMC 접근법을 적용하였다. 또한, 이 접근법은 Gibbs sampler 안에서 shrinkage를 위한 tuning parameter를 자동적으로 업데이트한다. 시뮬레이션 스터디는 본 논문의 방법론이 얼마나 효과적으로 (무시할 수 없는 간헐적인 missing data와 dropout에 의해 나타나는) estimation bias를 제거하는지를 나타낸다.

본 논문의 접근법은 outcome variables에 어떠한 분포가정도 하지 않기 때문에, 전통적인 평균회귀분석 (mean regression)보다 더 robust하다는 장점이 있다. 하지만 이는 dropput process에 대한 모델의 가정이 필요하다. 무시할 수 없는 missing data의 체계는 관찰된 데이터에 기반하여 직접적으로 식별할 수 없기 때문이다.

References

Yuan, Y., & Yin, G. (2009). Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data. Biometrics, 66(1), 105–114.