2019-1 | 베이지안통계특론 I

FINAL PROJECT

베이지안 분위 회귀모형(BQR)을 이용한 서울시 골목상권의 외식업종 매출 예측

학번	182STG18
이름	이하경
제출일	2019.06.11

DATA DESCRIPTION

▶ 분석 목표 정의

서울시는 '우리마을가게 상권분석서비스'를 통해 공공데이터를 개방하여 서울시에 분포한 상권영역별로 생활밀착업종(외식업, 서비스업, 도소매업 등)에 대해 정기적으로 추정매출과 상권별 인구 및 소득·소비에 대한 자료를 제공한다. 서울시에는 약 1000개 이상의 골목상권이 존재하고, 상권별로 평균 매출에 영향을 미치는 요인은 다양하다. 따라서 서울시에서 제공하는 상권 정보를 활용해 상권영역 및 업종별로 점포당 월평균매출액을 추정하고 매출액의 영향요인 및 의미에 대해 해석을 해보고자 한다. 골목상권의 위치와 업종에 따라 평균매출액의 규모에 차이가 있으므로, 조건부 평균에만 초점을 맞춘 평균 회귀모형보다 매출액의 전체적인 분포에 대한 변수별 영향력을 확인할 수 있는 **분위 회귀모형**(Quantile Regression)을 적용해보고자 한다.

서울열린데이터광장(data.seoul.go.kr)에서 분기별로 제공하는 서울시 골목상권 영역 내의 외식업종(한·중·일·양식음식점, 분식전문점, 치킨전문점, 패스트푸드점, 제과점, 커피·음료, 호프·간이주점)별 추정매출과 2018년 10월까지 월별로 제공된 상권영역별 추정 유동인구, 상주인구, 추정소득소비 자료를 사용하였다. 분기별 추정매출과 기타 자료를 함께 사용하기 위해 월별로 제공된 인구 및 소득소비에 관한 자료를 분기별로 평균해 총 2018년 1분기부터 3분기의 자료에 대해 분석을 진행하였다. 수집된 총 관측치는 8412개로, 분석에는 이중 1000개의 표본을 랜덤으로 추출하여 사용하였다.

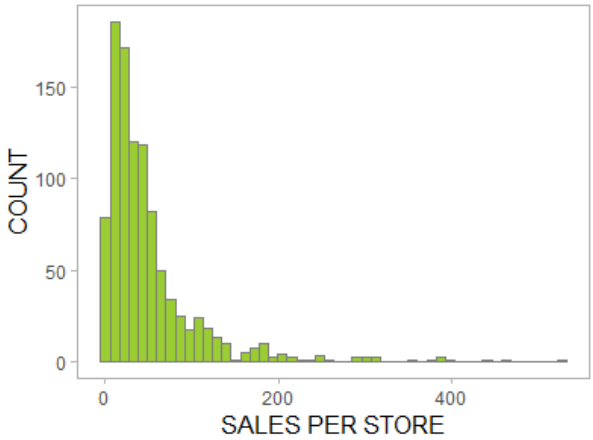
반응변수는 상권영역·외식업종별 점포당 월평균매출액으로 정의하였고, 설명변수는 다음과 같이 성별, 시간대, 연령이 각각 매출액에서 차지하는 비율과 골목상권별 해당 업종의 밀도를 나타내는 점포 수, 상권영역별 평균 유동인구와 상주인구, 월 평균 소득과 총 지출 금액으로 총 14개이다.

▶ 반응변수와 설명변수 정의

구분	변수		단위	비고
상권별 추정매출 (분기)	**Y**	**점포당 월 평균 매출 금액**	100만원	업종별 월평균매출액 / 점포 수
	X1	주말 매출 비율	%	
	X2	11~14시 매출 비율		
	X3	14~17시 매출 비율		
	X4	17~21시 매출 비율		
	X5	21시~24시 매출 비율		
	X6	여성 매출 비율		
	X7	10대 매출 비율		
	X8	20대 매출 비율		
	X9	30대 매출 비율		
	X10	점포 수	개	
상권별 유동인구(월)	X11	총 유동인구 수	명	분기별 평균 사용
상권별 상주인구(월)	X12	총 상주인구 수	명	분기별 평균 사용
상권별 소득소비(월)	X13	월 평균 소득 금액	원	분기별 평균 사용
	X14	총 소비(지출) 금액	원	분기별 평균 사용

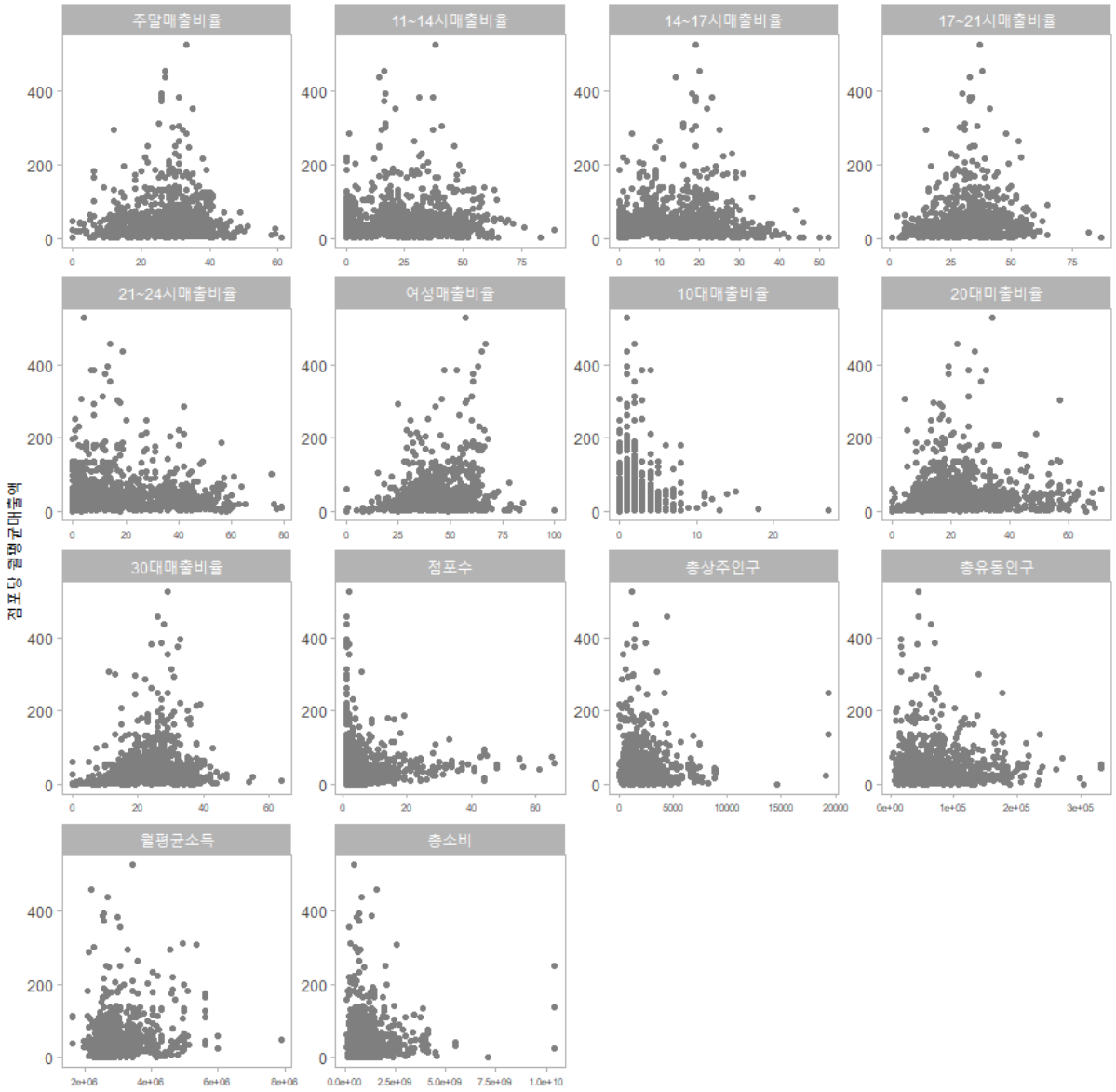
▶ 반응변수의 히스토그램과 기초통계량

Min	Q1	Mean	Q2	Q3	Max
0.099	15.166	32.372	48.818	58.037	526.705



반응변수 Y(점포당 월평균매출액, 단위: 100만원)의 히스토그램을 그려본 결과 꼬리가 오른쪽으로 긴 비대칭적인 분포의 모양을 하고 있음을 확인하였다.

▶ 반응변수와 설명변수의 산점도



위의 각 설명변수 X와 반응변수의 산점도에서 X의 값에 따라 Y의 분산이 일정하지 않은 것을 확인할 수 있다. 따라서 Y의 등분산성을 가정하고 주어진 X에 대한 Y의 평균에만 초점을 둔 평균 회귀분석은 사용이 적절하지 않으며 분위 회귀모형의 적용이 유용할 것으로 예상하였다.

METHODOLOGY

일반적으로 사용되는 평균 회귀모형은 반응변수 Y의 평균이 $x'\beta$인 정규분포를 따르며 x의 값에 상관없이 Y의 분산이 일정하다는 등분산성을 가정한다. 그러나 많은 경우의 실제 자료에서는 이 가정이 성립하지 않는다. 분위 회귀모형은 Y의 p분위수를 x의 값과 연결하여 p가 변화함에 따른 Y의 전체 분포에 대해서 설명변수의 영향을 탐색할 수 있다. 또한 p의 값에 따라 영향을 미치는 설명변수들이 서로 다를 수 있다. 분위 회귀모형은 Y의 정규성, 등분산성 등을 가정하지 않으므로 이상치의 영향을 크게 받지 않으며 보다 다양한 자료에 적합 가능하다.

베이지안 추론을 적용한 분위 회귀분석을 하기 위해서는 오차항에 대한 비대칭 라플라스 분포 가정으로 모형을 정의하고, 우도함수와 모수 β_p의 사전분포를 이용해 모수의 사후분포와 추정치를 도출한다.

$$Y = x\beta_p + \epsilon, \qquad \epsilon \sim ALD(p), \qquad \text{likelihood: } l(\beta_p|x,y) = \prod_{i=1}^{n} f_p(y_i, \beta, x_i) = p^n(1-p)^n \cdot e^{-\sum_{i=1}^{n}\rho_p(y_i - x_i\beta_p)}$$

본 분석 과제에서는 상권·업종별 점포당 평균매출액을 Y로 정의한 분위 회귀모형에 베이지안 추론을 적용하기 위해서 JAGS를 활용해 **p=(0.1, 0.25, 0.5, 0.75, 0.9)**의 다섯가지 분위에서 각각 절편을 포함한 **K=15**개의 분위 회귀계수 β_p의 사후표본을 생성하고 평균 및 편차와 신뢰구간을 구해보았다. 모형 설정에는 우도함수를 직접 지정하는 방법(**0-1 기법**)을 사용하였고, **깁스변수선택(GVS) 기법**을 사용해 평균매출액에 유의한 영향을 미치는 설명변수를 선별하기 위해 γ의 사전분포를 베르누이 분포로 설정하여 추가하였다. 회귀계수 β의 prior는 각각 독립적으로 N(0, 100)을 지정하였고, pseudo-prior는 R의 'quantreg' 패키지를 사용하여 얻은 평균과 분산의 추정치를 사용해 지정하였다.

모형에 입력할 설명변수들은 모두 사전에 평균을 0, 분산을 1으로 표준화하여 사용하였다.

또한 설정한 JAGS 모형에서 사전 설정한 다중 체인의 수 및 반복횟수는 각각 다음과 같다.

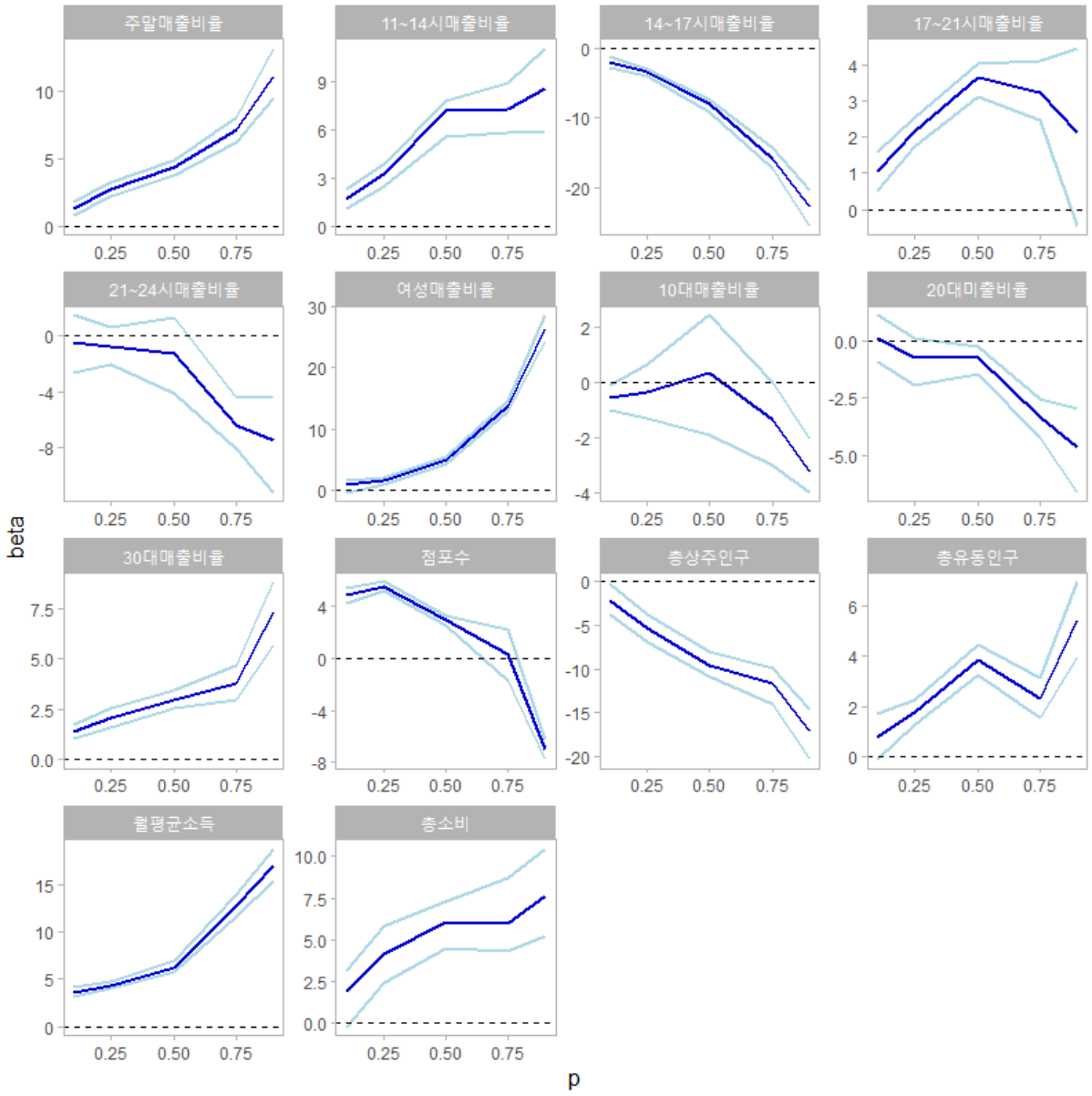
nChains	nAdapt	nUpdate	nIter	thin
3	10000	10000	20000	10

따라서 분위수 p마다 다중체인에서 얻어지는 각 β의 사후표본은 6000개이다. thining의 간격을 10으로 지정하여 총 반복횟수 20000번 중 10개의 표본마다 하나씩을 저장하고 추출되는 회귀계수 표본들 사이의 자기상관성을 줄이도록 하였다.

이렇게 얻어진 표본에서 γ의 조합 중 가장 상대빈도가 1위로 가장 높은 조합을 기준으로 설명변수들을 동시선택하고 해당 조합으로부터 산출된 표본만을 사용하여 최종적으로 회귀계수 β_p에 대한 사후 추정치 및 표준편차, 신뢰구간을 구하였다.

RESULT

[RESULT 1] 베이지안 분위회귀분석 결과: 분위에 따른 회귀계수 비교



위의 그림은 절편을 제외한 14개의 설명변수에 대해 분위수 p에 따라 추정된 분위회귀계수와 95% 신뢰구간을 나타낸다. 모든 변수에서 회귀계수의 추정치 및 유의성이 분위수 p에 대해 매우 달라짐을 확인할 수 있다.

- 주말 매출 비율, 여성 매출비율, 30대 매출 비율, 상권영역의 월평균 소득, 총소비는 더 높은 분위 모형일수록 양의 영향을 더 강하게 미친다. 이 변수들은 값이 클수록 점포당 매출액의 규모가 크다고 할 수 있다.

- 14-17시의 오후 시간대 매출 비율과 20대 매출비율, 총 상주인구는 낮은 분위에서는 영향이 작거나 유의하지 않지만 높은 분위일수록 음의 영향을 더 강하게 미친다. 이 변수들은 값이 클수록 점포당 매출액이 더 적다고 할 수 있다. 이것은 오후 시간대의 매출 비율이 클수록 음식점의 대부분의 매출을 차지하는 점심과 저녁 시간대의 매출 비율이 줄어드므로 적절한 해석이라고 할 수 있다.

- 점포 수의 경우 하위 분위에서는 양의 영향을 미치지만 0.75분위에서는 유의하지 않으며 더 높은 분위에서는 음의 영향을 미치는데, 매출액의 규모가 어느 정도 작은 분위까지는 점포 수가 증가할수록 점포당 매출액 역시 함께 증가하지만 매출액의 규모가 매우 큰 분위에서는 점포 수가 많을수록(또한 상주인구 또한 많을수록) 해당 상권영역의 밀집도가 높다고 할 수 있으므로 점포당 평균 매출액은 줄어든다는 해석이 가능하다.

[RESULT 2] 변수선택 결과 비교: stepAIC와 GVS

	p	주말	11·14시	14·17시	17·21시	21·24시	여성	10대	20대	30대	점포수	상주인구	유동인구	평균소득	총지출
GVS 동시선택 사후평균	0.1	●	●	●	●		●			●	●	●		●	●
	0.25	●	●	●	●		●		●	●	●	●	●	●	●
	0.5	●	●	●	●		●		●	●	●	●	●	●	●
	0.75	●	●	●	●	●	●	●	●	●		●	●	●	●
	0.9	●	●	●	●	●	●	●	●	●	●	●	●	●	●
stepAIC		●	●	●		●	●		●	●		●		●	●

- 각 분위수에서 최종 변수의 조합이 총 표본에서 차지하는 상대빈도는 각각 28.8%, 53.4%, 62.2%, 80.8%, 92.2%로 상위 분위수일수록 최적 모형이 더 확고하게 나타났다. γ의 전체 표본에서 변수별로 각각 계산된 사후평균을 기준으로 독립적으로 선택하는 방법 역시 동시선택과 동일한 결과를 보였다.

- 주말 매출 비율, 11~14시, 14~17시, 17~21시 매출 비율, 여성 매출 비율, 30대 매출 비율, 상권별 상주인구 및 월평균소득과 총지출금액 변수는 모든 분위수의 회귀모형에서 모두 최종 변수로 선택되었다. 상위 분위수 모형일수록 더 많은 변수들이 선택되어, 더 다양한 변수들이 유의한 영향을 미친다고 할 수 있다.

- p=0.5인 분위 회귀모형과 단계적 변수선택(stepwise)를 이용한 평균 회귀모형에서 최종선택된 변수를 비교할 때 분위 회귀모형에서는 평균 회귀모형과 달리 17시~21시의 매출 비율, 점포수, 유동인구가 추가로 유의한 변수로 나타났다.

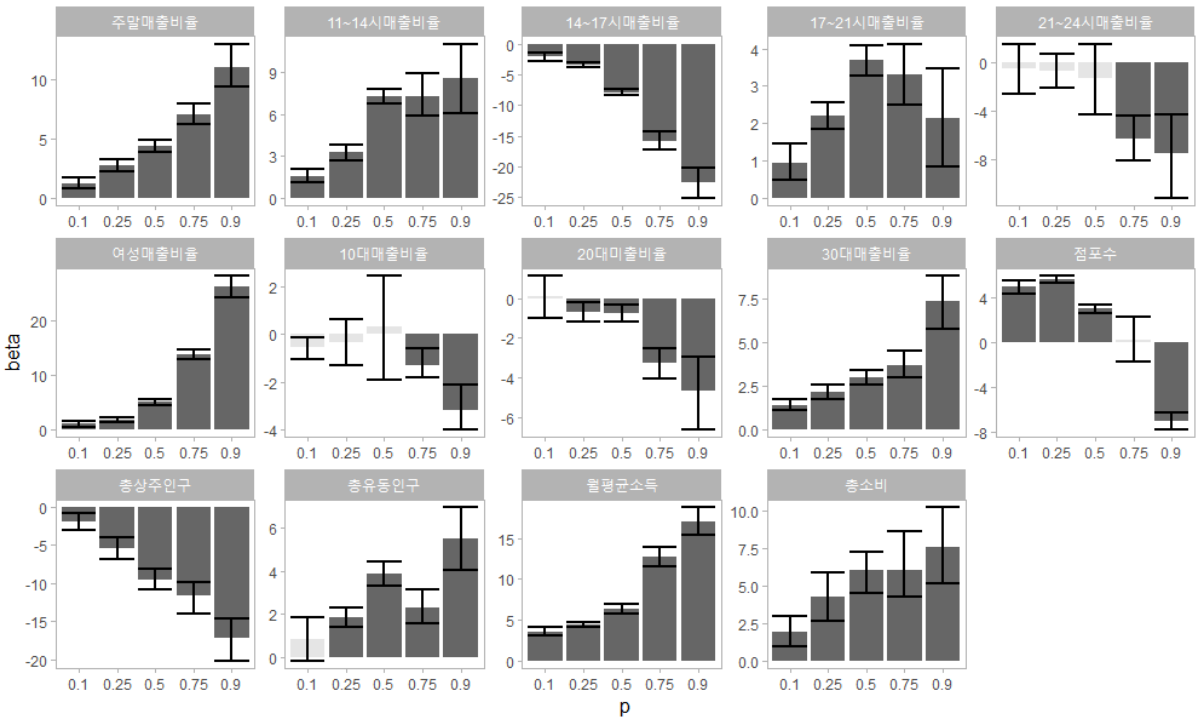
[RESULT 3] 최종 선택된 변수 별 회귀계수의 추론: 사후평균 및 표준편차, 신뢰구간

- 사후평균 및 사후표준편차

	p=0.1		p=0.25		p=0.5		p=0.75		p=0.9	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
INTERCEPT	10.438	0.240	18.047	0.224	33.319	0.232	58.191	0.388	103.853	0.663
주말	1.220	0.247	2.765	0.257	4.377	0.258	7.039	0.455	11.111	0.927
11-14시	1.568	0.266	3.274	0.295	7.254	0.267	7.305	0.770	8.582	1.290
14-17시	-2.066	0.331	-3.352	0.226	-7.882	0.259	-15.833	0.763	-22.613	1.233
17-21시	0.937	0.243	2.202	0.186	3.683	0.198	3.292	0.414	2.135	0.643
21시-24시	-	-	-	-	-	-	-6.309	0.926	-7.523	1.748
여성	0.934	0.298	1.709	0.273	5.050	0.275	13.825	0.480	26.296	0.998
10대	-	-	-	-	-	-	-1.321	0.337	-3.194	0.488
20대	-	-	-0.724	0.243	-0.748	0.219	-3.240	0.384	-4.649	0.932
30대	1.400	0.166	2.154	0.217	2.958	0.223	3.700	0.387	7.392	0.793
점포수	4.893	0.290	5.583	0.164	2.946	0.185	-	-	-7.018	0.367
상주인구	-1.928	0.527	-5.433	0.772	-9.626	0.685	-11.680	1.059	-17.217	1.395
유동인구	-	-	1.832	0.229	3.892	0.293	2.317	0.398	5.515	0.747
평균소득	3.530	0.261	4.394	0.155	6.275	0.307	12.687	0.618	17.099	0.875
총지출	1.956	0.522	4.237	0.827	6.033	0.687	6.064	1.133	7.624	1.265

- 사후 신뢰구간

	p=0.1		p=0.25		p=0.5		p=0.75		p=0.9	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
INTERCEPT	9.973	10.900	17.612	18.472	32.870	33.780	57.409	58.941	102.529	105.106
주말	0.746	1.690	2.260	3.259	3.840	4.868	6.193	7.953	9.431	13.055
11-14시	1.064	2.083	2.701	3.834	6.743	7.771	5.856	8.908	6.045	11.043
14-17시	-2.752	-1.448	-3.804	-2.907	-8.387	-7.378	-17.250	-14.270	-25.108	-20.292
17-21시	0.472	1.438	1.838	2.568	3.286	4.073	2.506	4.125	0.846	3.466
21시-24시	-	-	-	-	-	-	-8.022	-4.404	-11.106	-4.328
여성	0.346	1.513	1.145	2.219	4.496	5.581	12.869	14.747	24.422	28.382
10대	-	-	-	-	-	-	-1.833	-0.568	-3.991	-2.108
20대	-	-	-1.194	-0.231	-1.180	-0.324	-4.054	-2.528	-6.560	-2.947
30대	1.091	1.716	1.727	2.576	2.518	3.389	2.949	4.488	5.789	8.837
점포수	4.280	5.478	5.268	5.921	2.600	3.321	-	-	-7.771	-6.320
상주인구	-2.979	-0.894	-6.912	-3.939	-10.886	-8.138	-14.003	-9.934	-20.157	-14.727
유동인구	-	-	1.392	2.289	3.304	4.462	1.559	3.129	4.061	6.990
평균소득	3.049	4.065	4.103	4.716	5.730	6.927	11.553	13.916	15.435	18.833
총지출	0.981	2.998	2.632	5.865	4.549	7.272	4.287	8.631	5.173	10.278



위의 결과는 각 분위에서 상대빈도(사후확률)을 최고로 가지는 변수들의 조합과 대응되는 표본들만을 따로 추출하여 선택된 변수들의 회귀계수의 추정치와 신뢰구간을 다시 구한 결과이다. 총 6000개의 표본에서 각 사후확률이 계산된 빈도수만큼의 더 적은 표본을 추출하여 계산하였기 때문에 보다 더 정확한 추정치를 얻기 위해서는 선택된 변수들 만을 사용하여 반복 수를 크게 설정하여 다시 표본 추출을 진행해볼 수 있다.

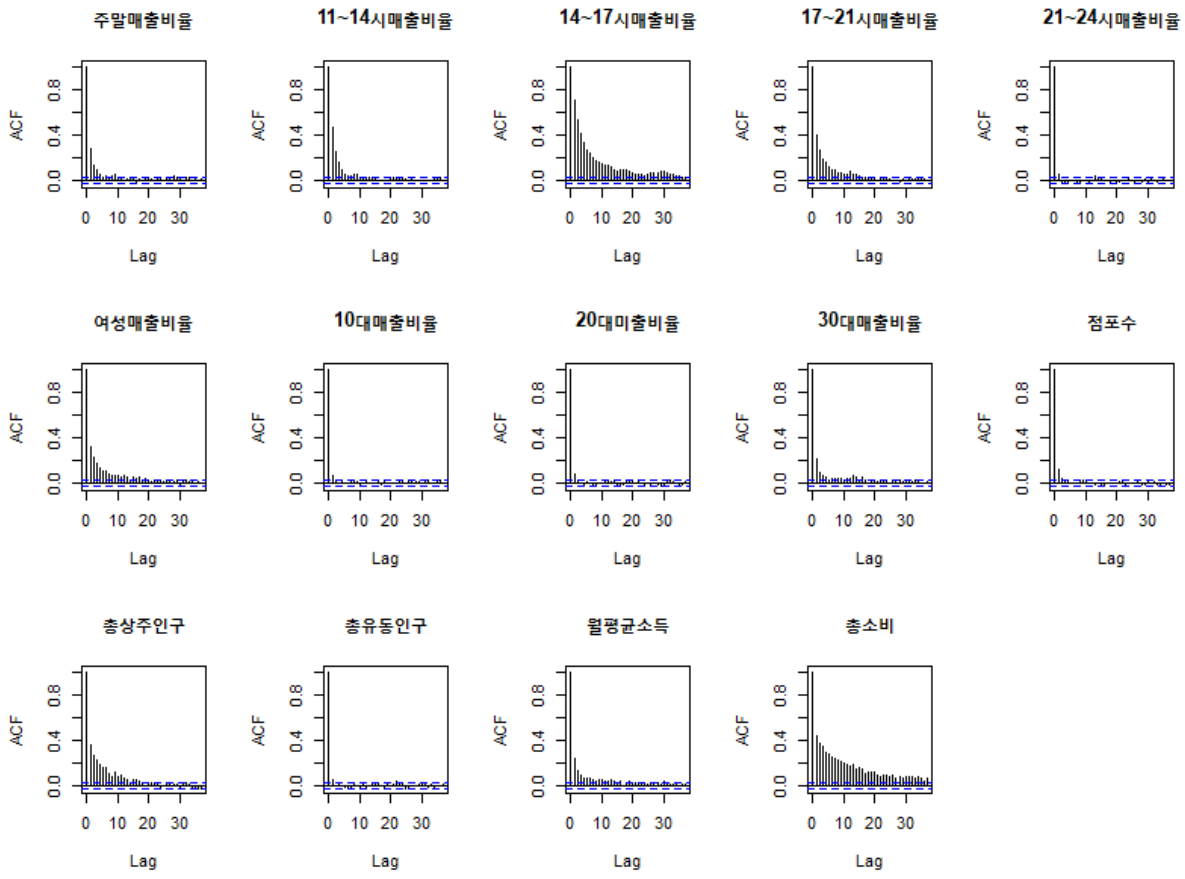
▶ 분위 회귀계수의 베이지안 사후 표본 겔만 상수 (Multivariate)

p=0.1	p=0.25	p=0.5	p=0.75	p=0.9
1.01	1.01	1.12	1.02	1.02

위의 표는 각 분위 회귀모형 당 저장된 6000개의 β_p 표본에 대하여 수렴성을 진단하기 위해 겔만 상수를 산출한 것이다. 아래에서 예시로 p=0.1의 자기상관함수와 추정된 사후밀도함수의 그림을 첨부하였으며, 다른 분위에 대한 그래프는 보고서에서 생략하였다. p=0.5 분위수에서 조금 더딘 점을 제외하고는 모두 MCMC를 이용한 회귀계수의 표본 추출이 적절히 이루어졌음을 확인하였다. 따라서 베이지안 표본 추출 및 사후 추론을 이용해 분위 회귀모형에서 다양한 설명변수의 값에 따른 반응변수 Y의 전체적인 분포를 추정할 수 있음을 알게 되었다.

더 나아가 이를 실제 서울시 골목상권 매출 자료에 적용해 상권영역 및 외식업종 별로 매출액의 규모에 따라 영향을 미치는 요인이 무엇인지 적절한 해석을 할 수 있었다. 주말과 여성, 30대 연령의 매출 비율이 클수록 모든 분위의 상권영역 점포당 매출액은 증가하며, 상위 분위에서 그 영향력이 더 커진다. 또한 상권영역의 평균 소득과 총소비 역시 동일한 경향을 보였다. 반면에 오후시간대와 20대 매출비율, 상권의 업종별 점포 수, 총 상주인구는 클수록 매출액이 작은 경향이 있으며 높은 분위에서 더 영향력이 큼을 확인하였다. 이처럼 반응변수의 여러 분위에서 설명변수들이 미치는 효과가 다양함을 직접 확인함으로써 평균 회귀분석보다 더 정밀한 추정이 가능함을 알게 되었다.

[첨부 1] (p=0.1) 분위 회귀모형 사후 표본의 자기상관함수



[첨부 2] (p=0.1) 분위 회귀모형 사후 표본의 사후밀도함수