



## 저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

이 학 석 사 학 위 논 문

질병지도 작성을 위해 공간모형을  
이용한 소지역 추정



부 경 대 학 교 대 학 원

통 계 학 과

안 대 성

이 학 석 사 학 위 논 문

질병지도 작성을 위해 공간모형을  
이용한 소지역 추정

지도교수 노 맹 석

이 논문을 석사 학위논문으로 제출함.

2015년 2월

부 경 대 학 교 대 학 원

통 계 학 과

안 대 성

안대성의 이학석사 학위논문을 인준함.

2015년 2월 27일



위원장	이학박사	장대흥 (인)
위원	이학박사	하일도 (인)
위원	이학박사	노맹석 (인)

# 목 차

표 차례	ii
그림 차례	ii
제 1장 서론 .....	1
제 2장 연구자료 및 방법 .....	4
2.1 연구자료 .....	4
2.2 연구방법 .....	11
2.2.1 분석모형 .....	11
2.2.2 모수추정 .....	13
2.2.3 변량효과에 대한 추론을 통한 질병지도 작성 .....	15
제 3장 연구결과 .....	17
3.1 질병 사망 .....	18
3.2 사고 사망 .....	20
3.3 암 사망 .....	22
3.4 뇌혈관질환 사망 .....	24
3.5 심장질환 사망 .....	26
제 4장 결론 및 제언 .....	28
참고문헌 .....	29
부록 .....	32

## 표 차례

1. Table 2.1 박탈지수 구성지표 .....	5
2. Table 3.1 질병 사망에 따른 모수 추정값 .....	18
3. Table 3.2 사고 사망에 따른 모수 추정값 .....	20
4. Table 3.3 암 사망에 따른 모수 추정값 .....	22
5. Table 3.4 뇌혈관질환 사망에 따른 모수 추정값 .....	24
6. Table 3.5 심장질환 사망에 따른 모수 추정값 .....	26

## 그림 차례

1. Figure 2.1. 2005년 인구센서스 기반 서울의 동별 박탈지수 .....	5
2. Figure 2.2. 질병 사망과 박탈지수와의 산점도 .....	6
3. Figure 2.3. 사고 사망과 박탈지수와의 산점도 .....	7
4. Figure 2.4. 암 사망과 박탈지수와의 산점도 .....	8
5. Figure 2.5. 뇌혈관질환 사망과 박탈지수와의 산점도 .....	9
6. Figure 2.6. 심장질환 사망과 박탈지수와의 산점도 .....	10
7. Figure 3.1 질병사망에 대한 MRF모형 적합 후 SMR 추정치 .....	19
8. Figure 3.2 사고사망에 대한 MRF모형 적합 후 SMR 추정치 .....	21
9. Figure 3.3 암사망에 대한 MRF모형 적합 후 SMR 추정치 .....	23
10. Figure 3.4 뇌혈관질환 사망에 대한 MRF모형 적합 후 SMR 추정치 ..	25
11. Figure 3.5 심장질환 사망에 대한 MRF모형 적합 후 SMR 추정치 .....	27

# Small Area Estimations for Disease Mapping by using Spatial Model

Dae Seong An

Department of Statistics, The Graduate School,  
Pukyong National University

## Abstract

SMRs (standardized mortality rates) for major diseases, accidents, cancer, cerebrovascular diseases, and cardiovascular diseases are considered in small areas of administrative units such as Eup/Myeon/Dong from years 2005 to 2008. Due to small sample issue in small areas, the precision of directly estimated crude SMR for each area can be poor. In this study, we consider the HGLM (hierarchical generalized linear model) with MRF (markov random field) to account for the spatial correlations among the small areas. The effects of covariates for cause of mortality by Dongs in Seoul and disease maps based on the estimated SMR are presented. The results suggest how we analyze and interpret the difference in mortalities by small areas such as Dongs by revealing the spatial patterns.

Keywords: disease mapping, hierarchical generalized linear model, small area estimation, spatially correlated model.

# 1. 서론

역학 (epidemiology), 보건행정 등 많은 분야에서는 행정구역 상 어떤 질병에 대한 위험(risk)이 지리적으로 분포하고 있는지가 주된 관심사일 수 있는데, 특히, 읍/면/동 소지역 (small area) 단위의 분석이 필요할 경우가 많다. 이를 위해, 각 소지역 행정단위별로 보고된 질병사망 자료를 가지고, 소지역 단위로 질병사망률을 직접 추정할 수 있다. 그러나, 소지역 내 표본수가 작아서 개별 소지역 단위의 직접 사망률 추정 결과는 그 추정치의 정도 (precision) 확보가 어려운 문제점으로 발생한다. 즉, 특정 소지역에서 특정 질병사건의 일부 발생/미발생으로 인하여 결과의 왜곡이 심해질 수 있다 (Kim과 Kim, 2009; Kim과 Sung, 2000; Park과 Lee, 2001; Rao, 2003).

이러한 문제점을 해결하기 위해 소지역 추정을 통한 질병지도 작성에 관한 통계적인 방법으로 Clayton과 Kaldor (1987), Ghosh 등 (1998)이 각각 제안한 경험적 베이즈 (empirical Bayes; EB) 및 계층적 베이즈 (hierarchical Bayes; HB) 추정방법들이 사용되어 왔다. 한편, 하나의 대안으로 Lee와 Nelder (1996, 2001)가 제안한 다단계 일반화 선형모형 (HGLM; hierarchical generalized linear model) 방법을 고려할 수 있다. 베이즈 접근법 (Banerjee 등, 2004)은 모수에 대한 사전분포 (prior distribution)를 가정하는 반면, HGLM 접근법은 기존의 우도 (likelihood)를 확장한 다단계 우도 (h-likelihood; hierarchical likelihood)에 기반하여 통계적 추론 방법을 제시하기 때문에, 사전분포 가정이 틀렸을 경우에 나타나는 모수추정치 민감성 (sensitivity) 문제가 발생

하지 않는다. 또한, 베イズ 접근법은 모수의 추정을 위해서는 MCMC (Markov chain Monte Carlo)와 같은 복잡한 계산과정을 거쳐야 하나, HGLM은 이런 복잡한 과정을 수행할 필요가 없고, MCMC보다 간단한 알고리즘을 통해 손쉽게 수행될 수 있다는 점에서 상당히 효과적인 분석 방법이 될 수 있다 (Kim 등, 2011; Lee, 등, 2006).

소지역간 효과를 인접 지역간 공간적 상관성이 있는 모형 (spatially correlated model)을 따르는 변량효과 (random effect)로 둔다면, 개별 소지역의 적은 표본 수로 인한 추정치의 정도에 대한 문제점을 해결할 수 있다. 이를 위해, 베이지안에서는 인접지역간 상관성을 랜덤워크 (random effect) 형태로 모형화한 CAR (conditional autoregressive) 모형을 일반적으로 사용하고 있다. 본 연구에서는 CAR 모형의 확장된 형태로 공간적 상관계수를 가지는 MRF (Markov random field) 모형을 고려하였다. MRF 모형은 다단계 우도에 의한 접근법을 통해 구현된 R 패키지 spaMM (Rousset, 2014)과 dhglm (Noh 와 Lee, 2011)을 통해 분석할 수 있다. 베이지안 CAR 모형은 WinBUGS와 같은 패키지로 구할 수 있지만, 현재까지 MRF 모형을 적합할 수 있는 베이지안 방법은 없다.

본 연구에서는 질병과 사고로 인한 표준화 사망비 (SMR; standardized mortality rate)가 서울특별시 행정동별로 어떠한 차이를 보이는지에 대해 분석하기 위해서 공간적 상관성을 가지는 HGLM을 고려하였다. 이를 위한 연구자료로 2005년도 기준 서울특별시 411개의 행정동에 대해서 2005-2008년 4년간 수집된 질병, 사고에 대한 사망자료를 활용하였다. 질병 중에서 암, 뇌혈관질환, 심장질환에 대해 추가적으로 분석하

였다. 각 소지역 단위인 행정동별로 2005년 기준으로 박탈지수 (deprivation index; Townsend, 1987)를 공변량 (covariate), 소지역별 차이를 변량효과 (random effects)로 고려하였다. 공변량 효과 및 변량효과를 추정하는 통해서는, 각 소지역간 질병사망의 연관성을 질병지도 (disease mapping)의 형태로 나타내어 파악하고자 한다. 지역공간상에서 발생하는 사망자 수는 포아송 분포를 따른다고 가정하고, 각 소지역별 기대사망자수의 자연로그 값을 오프셋 (offset)으로 두면 변량효과를 추정치는 SMR 값을 의미한다.

2장에서는 연구 자료의 소개로 본 연구에 이용된 자료의 변수에 대한 설명과 자료의 생성과정에 대해 기술하고, 연구방법으로 HGLM 분석모형 제시와 모수 및 지역별 효과 추정방법에 대하여 설명하고자 한다. 3장에서 연구결과로 질병 사망, 사고 사망, 암 사망, 뇌혈관질환 사망, 심장질환 사망에 대해서 공간적 상관성을 가지는 HGLM 분석모형을 이용하여, 각 모형별 공변량 및 공간적 상관성의 효과에 대해서 알아보고 추정된 SMR을 근거로 서울 지역 동별로 질병지도를 나타내었다. 4장에서는 연구결과를 요약하고 그에 따른 결론을 도출한다. 마지막으로, 부록에서는 분석을 위한 데이터 설명과 R코드를 첨부한다.

## 2. 연구자료 및 방법

### 2.1 연구자료

본 연구에서 반응변수는 2005-2008년 4년간 수집된 주요 사인별 사망자 수 (질병사, 사고사, 암사망, 뇌혈관질환 사망, 심장질환 사망)이며, 설명변수는 박탈지수이다. 동별 소지역에 따른 분석에 의미있는 결과를 주기 위해서는 1년 사망자료는 그 정보가 너무 희박하여 적어도 4년동안 수집된 자료가 적절한 것으로 판단된다. 주요 사인별 사망자 수는 통계청의 협조를 통해 411개 동별로 제공을 받았으며, 박탈지수는 지역의 사회적, 경제적 결핍 수준을 종합적으로 나타내는 지표이다. 이에 해당되는 동별 박탈지수를 산출하기 위해서는 5년마다 시행되는 인구센서스 조사의 10% 표본자료가 필요하며, 2005-2008년 4년간 사망자 수에 해당되는 박탈지수를 산출하기 위한 자료원은 2005년 인구센서스 10% 표본자료이며, 통계청의 협조를 통해 제공받았다.

박탈지수는 Choi 등 (2011)의 연구에 근거하여 다음 Table 2.1과 같이 9개 구성 지표로 구성하였다. 박탈지수는 음의 값이 클수록 박탈수준이 낮음 (사회경제적 수준이 높음)을 의미하며, 양의 값이 클수록 박탈수준이 높음 (사회경제적 수준이 낮음)을 의미한다 (Yoon, 2003).

Table 2.1 박탈지수 구성지표

박탈지수 구성지표	
1인 가구비율	여성 가구주 가구 비율
자가용이 없는 가구 비율	고졸 미만 교육수준의 인구 비율
낙후된 주거환경의 가구 비율	노인인구 비율
아파트가 아닌 가구 비율	가구주 기준 낮은 사회계층 비율

Figure 2.1에서 보는 바와 같이, 2005년 인구센서스를 토대로 박탈지수를 지도로 표현한 결과 서울의 서초, 강남, 잠실 지역이 강북 등의 지역에 비하여 사회적, 경제적 수준이 높아 상대적으로 박탈지수가 낮음을 알 수 있다.

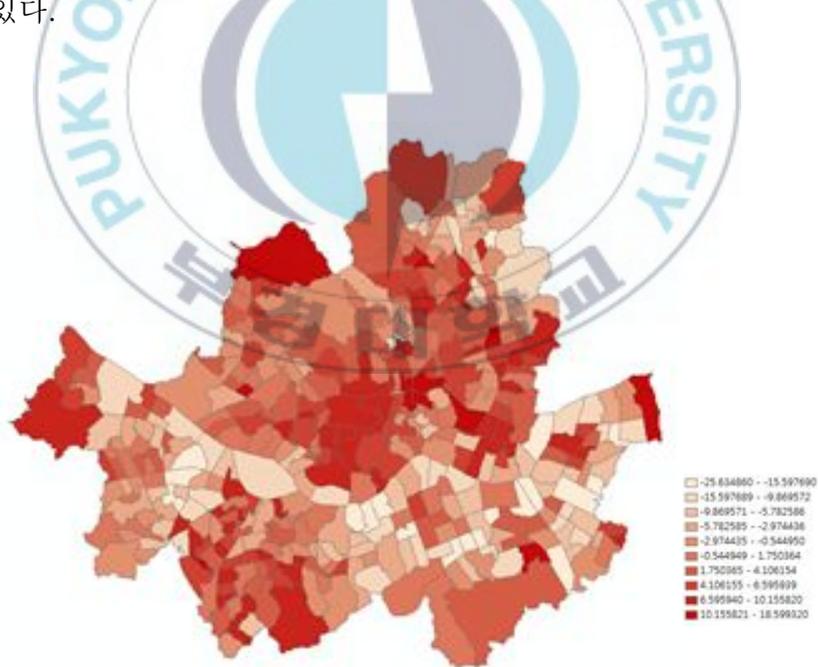


Figure 2.1. 2005년 인구센서스 기반 서울의 동별 박탈지수

Figure 2.2는 질병사망과 박탈지수와와의 산점도와 상관계수를 나타낸 것이다. 산점도를 보면 어느정도 양의 상관관계가 있을 것이라고 추정할 수 있다. 상관계수는  $r=0.562$  ( $p\text{-value} < 0.001$ )로 나타나 산점도에서 보는바와 같이 박탈지수와 의미있는 양의 상관관계를 보이고 있음을 알 수 있다. 즉, 사회적, 경제적 수준이 높은 지역일수록 질병 사망률이 낮음을 의미한다.

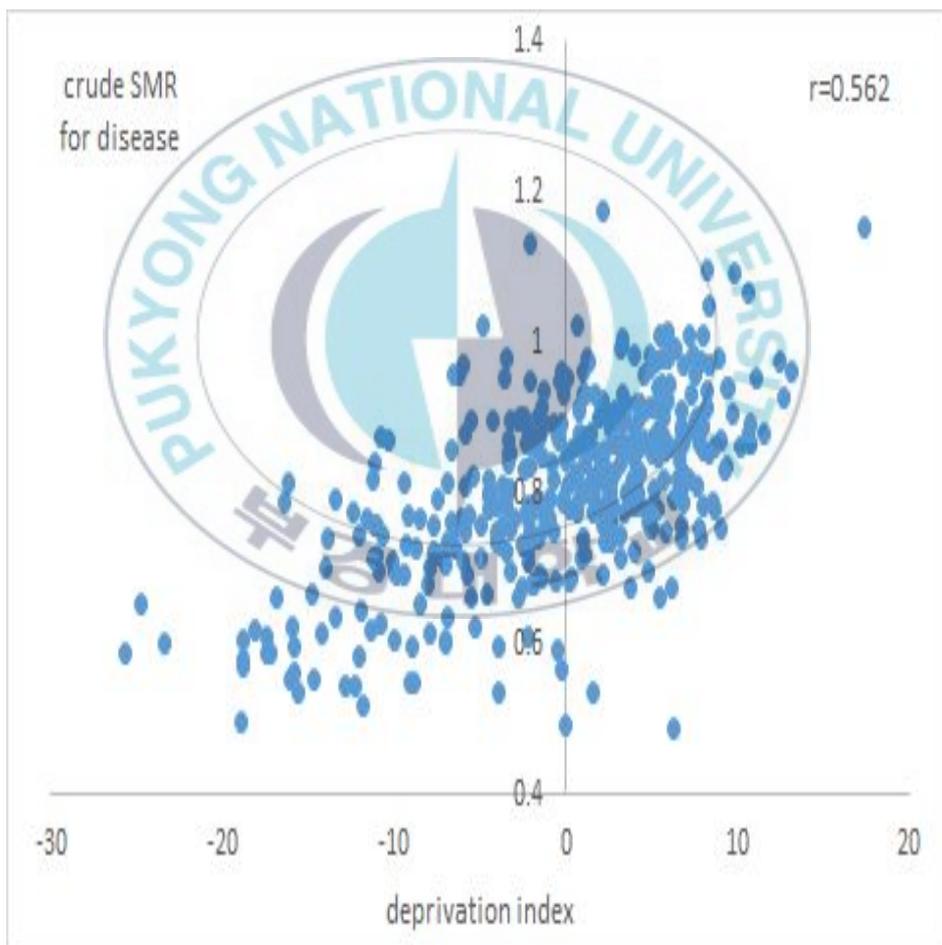


Figure 2.2. 질병사망과 박탈지수와와의 산점도

Figure 2.3은 사고사망과 박탈지수와의 산점도와 상관계수를 나타낸 것이다. 사고사망도 산점도에서 어느정도 양의 상관관계가 있을 것이라고 추정할 수 있다. 상관계수는  $r=0.559$  ( $p\text{-value} < 0.001$ )로 나타나 산점도에서 보는바와 같이 박탈지수와 의미있는 양의 상관관계를 보이고 있음을 알 수 있다. 즉, 사회적, 경제적 수준이 높은 지역일수록 사고 사망률이 낮음을 의미한다.

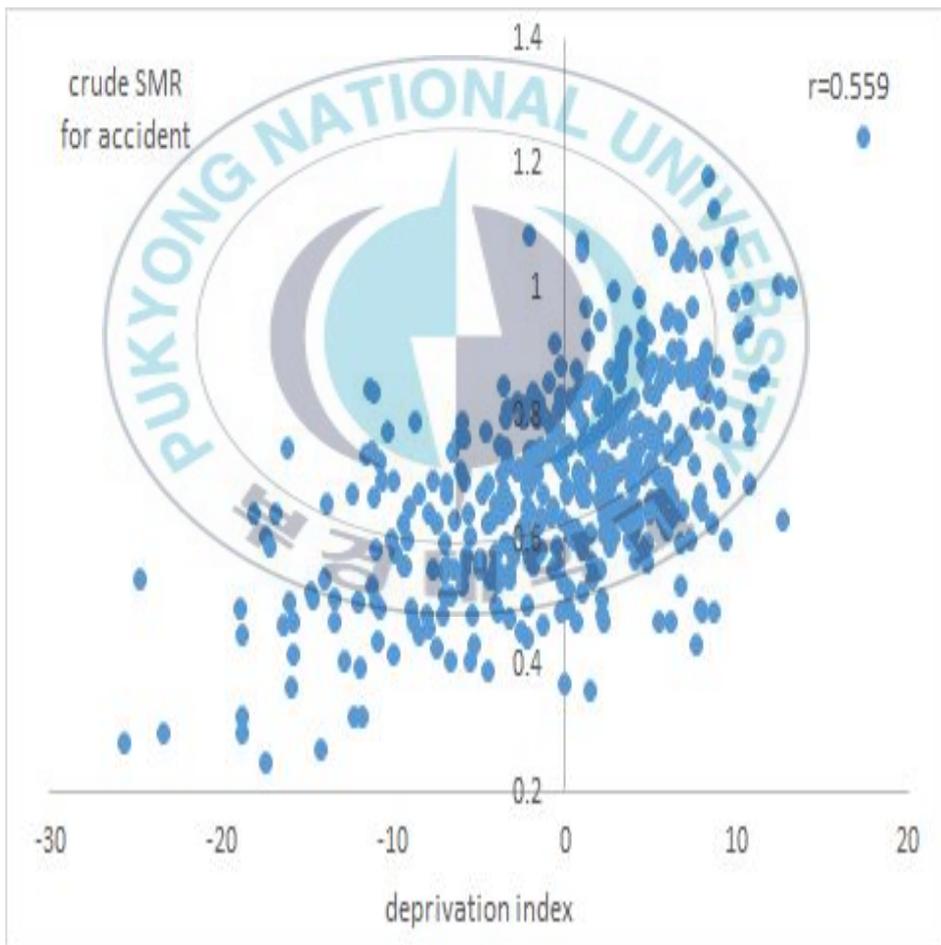


Figure 2.3. 사고사망과 박탈지수와의 산점도

Figure 2.4는 질병사망 중에서 암사망과 박탈지수와의 산점도와 상관 계수를 나타낸 것이다. 산점도에서 어느정도 양의 상관관계가 있을 것이라고 추정할 수 있다. 상관계수는  $r=0.378$  ( $p\text{-value} < 0.001$ )로 나타나 산점도에서 보는바와 같이 박탈지수와 의미있는 양의 상관관계를 보이고 있음을 알 수 있다. 즉, 사회적, 경제적 수준이 높은 지역일수록 암 사망률이 낮음을 의미한다.

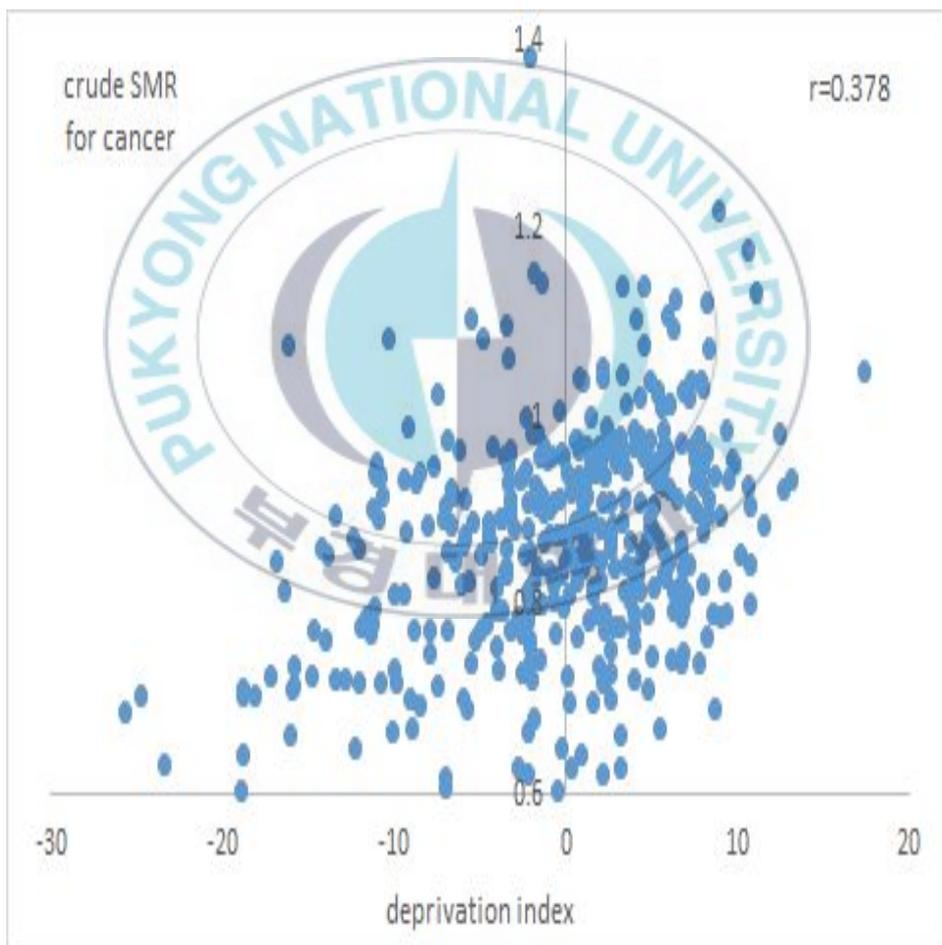


Figure 2.4. 암사망과 박탈지수와의 산점도

Figure 2.5는 질병사망 중에서 뇌혈관질환 사망과 박탈지수와의 산점도와 상관계수를 나타낸 것이다. 산점도에서 어느정도 양의 상관관계가 있을 것이라고 추정할 수 있다. 상관계수는  $r=0.392$  ( $p\text{-value} < 0.001$ )로 나타나 산점도에서 보는바와 같이 박탈지수와 의미있는 양의 상관관계를 보이고 있음을 알 수 있다. 즉, 사회적, 경제적 수준이 높은 지역일수록 뇌혈관질환 사망률이 낮음을 의미한다.

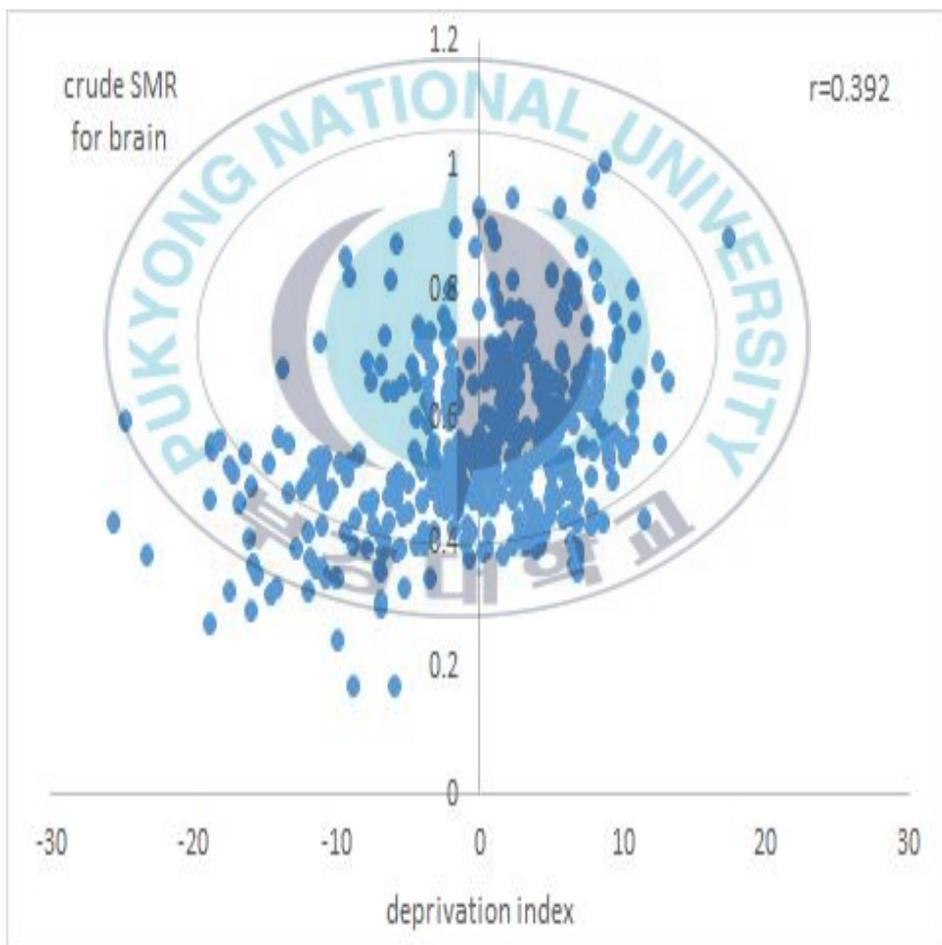


Figure 2.5. 뇌혈관질환 사망과 박탈지수와의 산점도

Figure 2.6은 질병사망 중에서 심장질환 사망과 박탈지수와의 산점도와 상관계수를 나타낸 것이다. 산점도에서 어느정도 양의 상관관계가 있을 것이라고 추정할 수 있다. 상관계수는  $r=0.392$  ( $p\text{-value} < 0.001$ )로 나타나 산점도에서 보는바와 같이 박탈지수와 의미있는 양의 상관관계를 보이고 있음을 알 수 있다. 즉, 사회적, 경제적 수준이 높은 지역일수록 심장질환 사망률이 낮음을 의미한다.

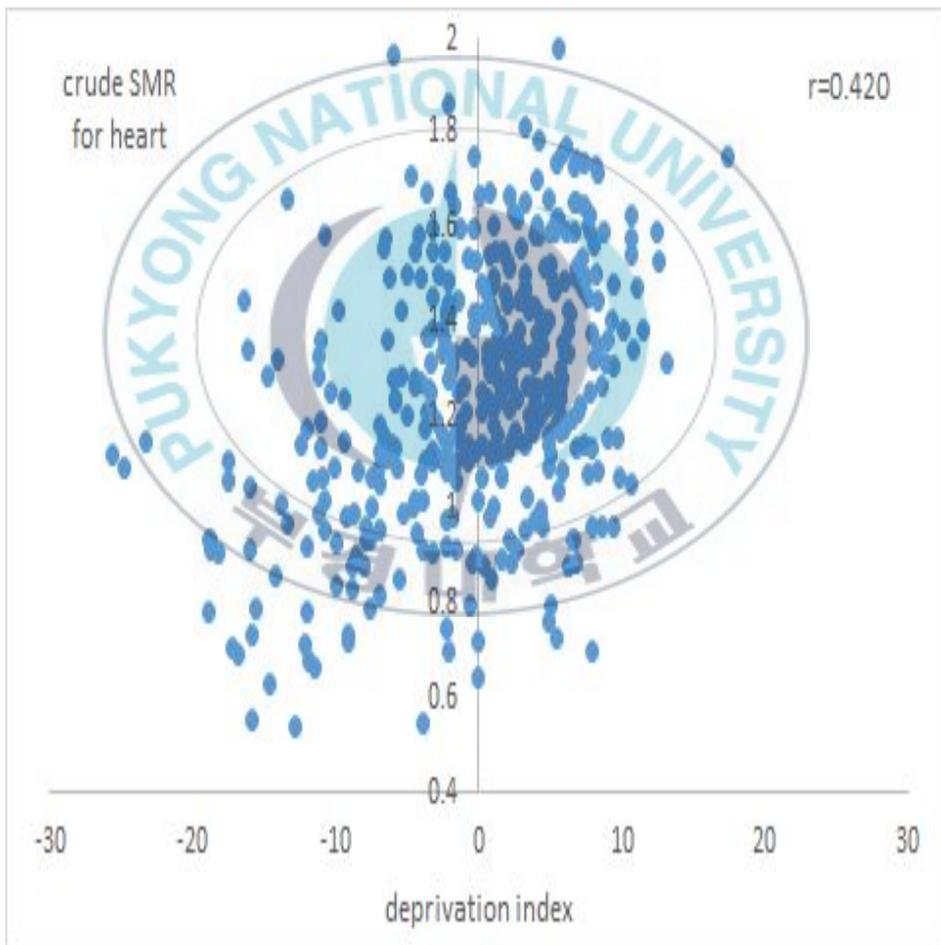


Figure 2.6. 심장질환 사망과 박탈지수와의 산점도

## 2.2 연구방법

본 장에서는 공간적 상관성이 있는 HGLM 모형과 이를 적합하기 위해 다단계 우도에 의한 방법을 소개하고자 한다.

### 2.2.1 분석모형

$y_i$ 를  $i$  ( $i=1, \dots, n=411$ )번째 소지역 단위인 2005년부터 2008년까지의 특정 사망원인 (질병, 사고, 암)에 대한 사망자 수라고 하였을 때, 식 (2.1)과 같은 모형을 고려할 수 있다. 즉, 소지역 효과  $v_i$ 가 주어졌을 때  $y_i$ 는 조건부 기대도수  $\mu_i$ 을 가지는 포아송 분포를 따른다고 가정한다.  $v_i$ 는 분포가정을 하기 때문에, 포아송 HGLM이 된다. 이러한 모형은 흔히 BYM (Besag 등, 1991) 모형이라고 부른다.  $\mu_i$ 에 대한 연결함수 (link function)를 로그함수로 둔 식 (2.2)와 같은 모형을 고려하였고, 이때  $\eta_i$ 는  $\mu_i$ 에 대한 선형 예측 (linear predictor) 변수가 된다.

$$y_i | v_i \sim \text{Poisson}(\mu_i) \quad (2.1)$$

$$\eta_i = \log \mu_i = \log(E_i) + \beta_0 + \beta_1 x_i + v_i \quad (2.2)$$

이때,  $y_i$ 는  $i$ 번째 소지역의 실제 사망자 수를 의미하며,  $x_i$ 는  $i$ 번째 소지역의 박탈지수를 의미한다.  $E_i$ 는  $i$ 번째 소지역의 인구수에 대한 서울 전체의 사망률을 적용시킨 기대사망자 수가 되며,  $\log(E_i)$ 는 오프셋 (offset)으로 모형화 한다.  $SMR_i$ 는 공변량인  $x_i$ 의 효과를 고정하였을 때,  $i$ 번째 동의 기대도수 대비 관측도수의 비율을 나타낸다. 만약  $SMR_i$ 이 1

보다 크면 기대사망자보다 그 지역의 사망자수가 많다는 것을 의미하며, 1보다 작으면 기대되는 사망자보다 그 지역의 사망자수가 작다는 것을 의미한다. 예를 들어,  $i$ 번째 동의  $SMR_i$ 이 1.05라면,  $x_i$ 의 효과를 고정하였을 때, 전체 서울지역의 인구수와  $i$ 번째 동의 인구수를 비교한 예측 사망보다 실제사망이 1.05배 높다고 할 수 있다. 이러한,  $SMR_i$ 는  $\exp(v_i)$ 를 추정함으로써 각 동별로 추정할 수 있다.

소지역 효과  $v = (v_1, \dots, v_n)^T$ 는 다변량 정규분포를 통해 다음과 같이 독립모형 (M1)과, 공간적 상관성을 고려한 MRF (Markov random field) 모형 (M2)를 고려할 수 있다.

$$M1: v_i \stackrel{iid}{\sim} N(0, \lambda) \quad (i=1, \dots, n),$$

공간적으로 독립인 모형, MRF에서  $\rho = 0$

$$M2: v \sim MRF, \quad \Sigma^{-1} = [\text{var}(v)]^{-1} = (1 - \rho N) / \lambda$$

여기에서,  $N$ 은 서울411개동에 대한 인접정보를 담고 있는  $411 \times 411$  행렬로서, 만약  $i$ 번째 동과  $j$ 번째 동이 인접하였으면 행렬  $N$ 의  $(i, j)$ 번째 원소는 1, 인접하지 않았으면 0을 가진다. MRF 모형에서  $\rho$ 는 공간적 상관성을 나타내는 상관계수로서 1에 가까울수록 각 동별 상관성이 높으며 0에 가까울수록 상관성이 없고 -1에 가까울수록 역상관의 관계를 가지는 모형을 의미한다.

따라서, MRF 모형에서  $\rho$ 가 0이면  $[\text{var}(v_i)]^{-1} = 1/\lambda$  되어, M1 모형

즉,  $v_i \stackrel{iid}{\sim} N(0, \lambda)$ 인 독립인 모형이 된다.

### 2.2.2 모수 추정

모형(2.1)을 적합하기 위해서, 반응변수  $y = (y_1, \dots, y_n)^T$  (단,  $n=411$ )에 대한 주변 로그-우도 (marginal log-likelihood)  $m$ 은 다음 식 (2.3)과 같이 정의된다.

$$m = \log f(y) = \log \left( \int \prod_{i=1}^n \exp(f(y_i|v_i)) f(v) dv \right) \quad (2.3)$$

여기서, 주변 분포  $f(y)$ 는 결합분포  $f(y, v)$ 에서 변량효과  $v$ 에 대해 적분한 형태이다. 만약,  $v = (v_1, \dots, v_n)^T$ 가 서로 독립인 M1과 같은 모형이 가정되었을 때는 식 (2.4)와 같은 형태가 되며, SAS의 NLMIXED 프로시저에서 제공하는 GHQ (Gauss-Hermite quadrature) 방법을 사용하여 모수를 추정할 수 있다.

$$m = \sum_{i=1}^n \log \left( \int \exp(f(y_i|v_i)) f(v_i) dv_i \right) \quad (2.4)$$

그러나, 인접지역간 공간적 상관성을 가지는 M2 모형에서는 식 (2.4)는 411차원의 적분이 필요하게 되기 때문에, GHQ 방법으로서의 적합은 도저히 불가능하며 베이지안 방법을 통한 방법 역시 적합이 쉽지 않다. 따라서, 다단계 우도에 의한 방법을 적용할 수 있는데, M2 모형에 대한

다단계 우도는 식 (2.5)와 같이 정의할 수 있다.

$$h = h(\beta_0, \beta_1, \rho, \lambda; \beta_0, \beta_1) = \sum_{i=1}^n \log f(y_i | v_i; \beta_0, \beta_1) + \log f(v) \quad (2.5)$$

$$\log f(y_i | v_i; \beta_0, \beta_1) = y_i \eta_i - \exp(\eta_i) - \log(y_i!),$$

$$\log f(v; \rho, \lambda) = -\frac{1}{2} v^T \Sigma^{-1} v - \frac{1}{2} \log |2\pi \Sigma|,$$

$$\text{단, } \Sigma^{-1} = (1 - \rho N) / \lambda$$

식 (2.5)에서 보는 바와 같이 다단계 우도는 적분없이 정의되기 때문에, 보다 간단히 모수를 추정할 수 있다. 변량효과  $v$ 는  $h$ 를, 평균 모수  $(\beta_0, \beta_1)$ 는 수정-단면 우도 (adjusted profile likelihood)인  $p_v(h)$ 를, 산포 모수인  $(\rho, \lambda)$ 는 제한 우도 (restricted likelihood)  $p_{\beta_0, \beta_1, v}(h)$ 를 최대 로 하는 값을 추정치로 구할 수 있다 (Lee 등, 2006). 이때, 수정-단면 우도  $p_v(h)$ 는 식 (2.6)과 같이 정의되며, 마찬가지로 방법으로  $p_{\beta_0, \beta_1, v}(h)$ 를 정의할 수 있다.

$$p_v(h) = \left\{ h - \frac{1}{2} \log |D(h, v) / 2\pi| \right\}_{v = \tilde{v}} \quad (2.6)$$

$$\text{단, } D(h, v) = -\partial^2 h / \partial v \partial v^t,$$

$$\tilde{v} : \partial h / \partial v = 0 \text{ 을 만족하는 해}$$

### 2.2.3 변량효과에 대한 추론을 통한 질병지도 작성

HGLM에서 변량효과 각  $v_i$  들의 추정치인  $\hat{v}_i$  는 다단계 우도  $h = \sum_{i=1}^n h_i$  을 최대로 하는 값을 통해서 구할 수 있으며, 적절한 조건하에서 점근적으로  $\hat{v}_i = E(v_i | y_i)$  가 된다.  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$  와  $\hat{v} - v$  의 점근적인 공분산 행렬 (asymptotic covariance matrix) 은 식 (2.7) 과 같은 Hessian 행렬인  $H$  의 역행렬로부터 얻어진다 (Lee와 Nelder, 1996).

$$H(\beta, v) = - \begin{pmatrix} \frac{\partial^2 h}{\partial \beta^2} & \frac{\partial^2 h}{\partial \beta \partial v} \\ \frac{\partial^2 h}{\partial v \partial \beta} & \frac{\partial^2 h}{\partial v^2} \end{pmatrix} = \begin{pmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + R \end{pmatrix} \quad (2.7)$$

여기서,  $X$  와  $Z$  는 각각  $\beta$ ,  $v$  에 대한 모형행렬 (model matrix) 이며,  $W = \text{diag}(\mu_{ij})$  는 가중치 대각행렬 (weighted diagonal matrix) 이며,  $R_{q \times q} = \text{diag}\{-\partial^2 \log(f(v_i; \sigma_v^2) / \partial v_i^2)\}$  이다. 이때,  $\hat{v} - v$  의 공분산행렬은 식 (2.7) 의  $H(\beta, v)$  의 역행렬에서 오른쪽 아래부분 (right-hand corner) 인 식 (2.8) 을 통해서 구할 수 있다.

$$\text{Var}(\hat{v} - v) = \{(Z^T W Z + R) - (Z^T W X)(X^T W X)^{-1}(X^T W Z)\}^{-1} \quad (2.8)$$

따라서, 지역별 변량효과  $v_i$  의 95% 신뢰구간은 식 (2.9)과 같으며, 이때 표준오차 (SE; standard error) 추정치는 변량효과  $v_i$ 에 대한 적절한 신뢰구간 추정치를 제공해 준다 (Lee와 Ha, 2010).

$$\hat{v}_i \pm 1.96 \times SE(\hat{v}_i - v_i) \quad (2.9)$$

단,  $SE(\hat{v}_i - v_i) = \sqrt{\text{var}(\hat{v}_i - v_i)}$



### 3. 연구결과

본 장에서는 연구결과로 질병사망, 사고사망, 암 사망, 뇌혈관질환 사망, 심장질환 사망에 대한 공변량의 효과 및 소지역별로 추정된 SMR을 근거로 모수추정값과 서울특별시 행정동별로 질병지도를 제시하고자 한다. 박탈지수는 공변량으로 고려하였을 때, 각 사망원인에 대한 M1, M2 적합결과를 Table 3.1~Table 3.5에 각각 제시 하였다. Figure 3.1~Figure 3.5는 각 사망원인별로 지역간 상관성을 고려한 M2모형에서 박탈지수를 공변량으로 고려하지 않은 모형 (상)과 고려한 모형 (하)을 적합한 후 SMR 추정치를 나타내고 있다. t-value는 추정치/표준오차 (estimate/SE)를 의미하며, 모수가 0이라는 귀무가설 하에서 점근적으로 표준정규분포를 따른다. 박탈지수에 대한 t-value가 1.96보다 크다면 박탈지수는 5% 유의수준 하에서 통계적으로 유의하게 SMR에 영향을 준다고 할 수 있다. M1, M2 모형 적합 결과에서 나오는 제한우도값  $p_{\beta_0, \beta_1, v}(h)$ 을 사용하여  $H_0: \rho=0$ 에 대한 가설검정을 수행할 수 있다.  $H_0: \rho=0$ 에 대해서 우도비 검정 (LRT; likelihood ratio test) 통계량은  $LR=-2(RL(H_0)-RL(H_1))$ 으로 정의되며,  $H_0: \rho=0$  하에서 자유도가 1인 카이제곱 분포를 따른다. 이때,  $RL(H_0)$ 와  $RL(H_1)$ 은 각각 귀무가설, 대립가설 하에서의 제한우도 값을 나타낸다.

### 3.1 질병 사망

Table 3.1에서 보는 바와 같이 질병사망에 대해서  $H_0 : \rho=0$ 에 대한 우도비 검정 결과 인접지역간 공간적 상관성이 아주 유의한 것으로 나타났다.  $\rho$  추정치는 질병사망은  $\hat{\rho}=0.150$ 로 나타났다. 공간적 상관성을 고려한 M2 모형에서 박탈지수는 양의 계수로 SMR에 유의하게 영향을 주고 있음을 알 수 있다. 즉, 사회·경제적인 수준이 높은 지역은 그렇지 않은 지역에 비하여 고려된 질병사망의 발생률이 낮음을 의미한다.

Table 3.1 질병사망에 따른 모수 추정값

parameter	M1 : 독립모형			M2 : MRF모형		
	estimate	SE	t-value	estimate	SE	t-value
Intercept	-0.211	0.00579	-36.46	-0.193	0.01213	-15.9
박탈지수	0.014	0.00083	16.69	0.013	0.00084	14.89
$\log\lambda$	-4.57	0.11		-5.02	0.12	
LRT				LR=74.6 ( $p < 0.001, H_0 : \rho = 0$ )		
$\rho$	0			0.150		

Figure 3.1에서 보는 바와 같이 박탈지수를 공변량으로 고려하지 않았을 때는 소지역인 동별로 질병사망에 대한 SMR 추정치가 뚜렷한 차이를 가지고 있는 반면에, 공변량으로 고려하면 이러한 차이가 사라지는 것을 알 수 있다. 즉, 박탈지수가 질병사망에 대한 SMR의 동별 차이를 잘 설명하고 있음을 의미한다.

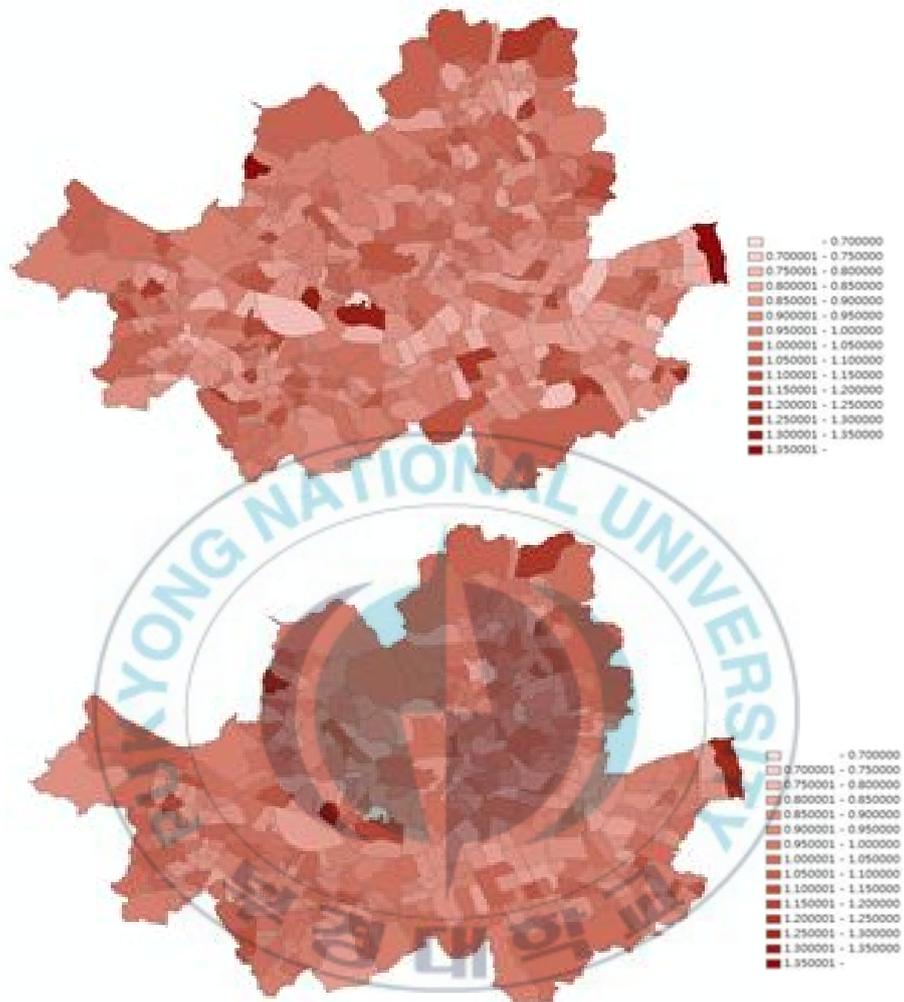


Figure 3.1 질병사망에 대한 MRF모형 적합 후 SMR 추정치

(상 : 공변량 없음, 하 : 공변량 고려)

### 3.2 사고 사망

Table 3.2에서 보는 바와 같이 사고 사망에 대해서  $H_0 : \rho = 0$ 에 대한 우도비 검정 결과 인접지역간 공간적 상관성이 아주 유의한 것으로 나타났다.  $\rho$  추정치는  $\hat{\rho} = 0.146$ 로 나타났다. 공간적 상관성을 고려한 M2 모형에서 박탈지수는 양의 계수로 SMR에 유의하게 영향을 주고 있음을 알 수 있다. 즉, 사회·경제적인 수준이 높은 지역은 그렇지 않은 지역에 비하여 고려된 사고사망의 발생률이 낮음을 의미한다.

Table 3.2 사고사망 따른 모수 추정값

parameter	M1 : 독립모형			M2 : MRF모형		
	estimate	SE	t-value	estimate	SE	t-value
Intercept	-0.375	0.00951	-39.39	-0.376	0.0159	-23.65
박탈지수	0.021	0.00142	14.81	0.02	0.00151	13.38
log $\lambda$	-4.35	0.17		-4.77	0.18	
LRT	LR=14.7 ( $p < 0.001, H_0 : \rho = 0$ )					
$\rho$	0			0.146		

Figure 3.2에서 보는 바와 같이 박탈지수를 공변량으로 고려하지 않았을 때는 소지역인 동별로 사고사망에 대한 SMR 추정치가 뚜렷한 차이를 가지고 있는 반면에, 공변량으로 고려하면 이러한 차이가 사라지는 것을 알 수 있다. 즉, 각 박탈지수가 사고사망에 대한 SMR의 동별 차이를 잘 설명하고 있음을 의미한다.

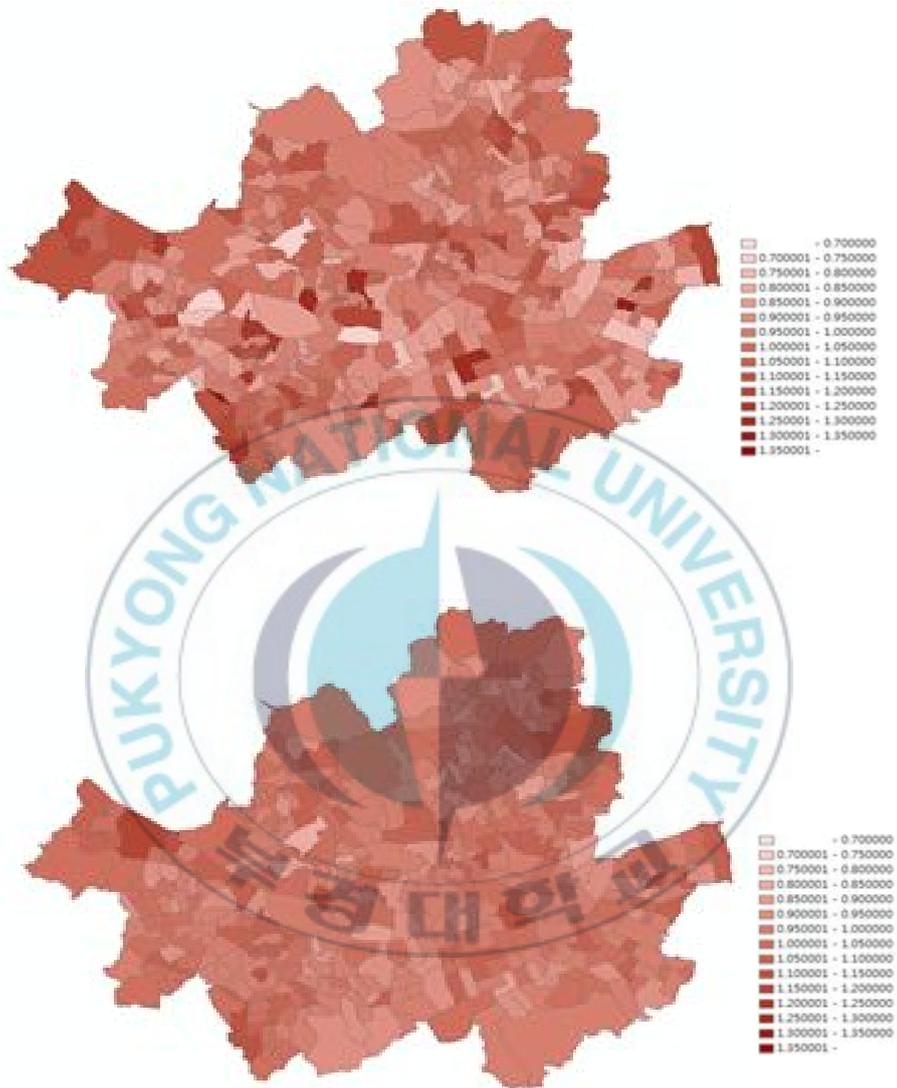


Figure 3.2 사고사망에 대한 MRF모형 적합 후 SMR 추정치

(상 : 공변량 없음, 하 : 공변량 고려)

### 3.3 암 사망

Table 3.3에서 보는 바와 같이 암 사망에 대해서  $H_0 : \rho=0$ 에 대한 우도비 검정 결과 인접지역간 공간적 상관성이 아주 유의한 것으로 나타났다.  $\rho$  추정치는  $\hat{\rho}=0.152$ 로 나타났다. 공간적 상관성을 고려한 M2 모형에서 박탈지수는 양의 계수로 SMR에 유의하게 영향을 주고 있음을 알 수 있다. 즉, 사회·경제적인 수준이 높은 지역은 그렇지 않은 지역에 비하여 고려된 암사망의 발생률이 낮음을 의미한다.

Table 3.3 암사망 따른 모수 추정값

parameter	M1 : 독립모형			M2 : MRF모형		
	estimate	SE	t-value	estimate	SE	t-value
Intercept	-0.14	0.0063	-22.43	-0.13	0.0121	-10.72
박탈지수	0.008	0.0009	8.7	0.007	0.00097	6.73
log $\lambda$	-5.03	0.15		-5.57	0.18	
LRT	LR=37.9 ( $p < 0.001, H_0 : \rho = 0$ )					
$\rho$	0			0.152		

Figure 3.3에서 보는 바와 같이 박탈지수를 공변량으로 고려하지 않았을 때는 소지역인 동별로 암 사망에 대한 SMR 추정치가 뚜렷한 차이를 가지고 있는 반면에, 공변량으로 고려하면 이러한 차이가 사라지는 것을 알 수 있다. 즉, 각 박탈지수가 암 사망에 대한 SMR의 동별 차이를 잘 설명하고 있음을 의미한다.



Figure 3.3 암사망에 대한 MRF모형 적합 후 SMR 추정치

(상 : 공변량 없음, 하 : 공변량 고려)

### 3.4 뇌혈관질환 사망

Table 3.4에서 보는 바와 같이 뇌혈관질환 사망에 대해서  $H_0 : \rho = 0$ 에 대한 우도비 검정 결과 인접지역간 공간적 상관성이 아주 유의한 것으로 나타났다.  $\rho$  추정치는  $\hat{\rho} = 0.154$ 로 나타났다. 공간적 상관성을 고려한 M2 모형에서 박탈지수는 양의 계수로 SMR에 유의하게 영향을 주고 있음을 알 수 있다. 즉, 사회·경제적인 수준이 높은 지역은 그렇지 않은 지역에 비하여 고려된 뇌혈관질환 사망의 발생률이 낮음을 의미한다.

Table 3.4 뇌혈관질환 사망에 따른 모수 추정값

parameter	M1 : 독립모형			M2 : MRF모형		
	estimate	SE	t-value	estimate	SE	t-value
Intercept	-0.5770	0.0107	-53.86	-0.580	0.0194	-29.91
박탈지수	0.0134	0.0016	8.48	0.013	0.0016	8.31
log $\lambda$	-4.19	0.17		-5.24	0.24	
LRT	LR=79.6 ( $p < 0.001, H_0 : \rho = 0$ )					
$\rho$	0			0.154		

Figure 3.4에서 보는 바와 같이 박탈지수를 공변량으로 고려하지 않았을 때는 소지역인 동별로 뇌혈관질환 사망에 대한 SMR 추정치가 뚜렷한 차이를 가지고 있는 반면에, 공변량으로 고려하면 이러한 차이가 사라지는 것을 알 수 있다. 즉, 각 동별 박탈지수가 뇌혈관질환 사망에 대한 SMR의 동별 차이를 잘 설명하고 있음을 의미한다.



Figure 3.4 뇌혈관질환 사망에 대한 MRF모형 적합 후 SMR 추정치

(상 : 공변량 없음, 하 : 공변량 고려)

### 3.5 심장질환 사망

Table 3.5에서 보는 바와 같이 심장질환 사망에 대해서  $H_0 : \rho=0$ 에 대한 우도비 검정 결과 인접지역간 공간적 상관성이 아주 유의한 것으로 나타났다.  $\rho$  추정치는  $\hat{\rho}=0.149$ 로 나타났다. 공간적 상관성을 고려한 M2 모형에서 박탈지수는 양의 계수로 SMR에 유의하게 영향을 주고 있음을 알 수 있다. 즉, 사회·경제적인 수준이 높은 지역은 그렇지 않은 지역에 비하여 고려된 심장질환 사망의 발생률이 낮음을 의미한다.

Table 3.5 심장질환 사망에 따른 모수 추정값

parameter	M1 : 독립모형			M2 : MRF모형		
	estimate	SE	t-value	estimate	SE	t-value
Intercept	0.226	0.0095	23.88	0.222	0.0171	12.96
박탈지수	0.013	0.0014	9.59	0.013	0.0015	8.51
log $\lambda$	-4.29	0.16		-4.75	0.179	
LRT	LR=22.0 ( $p < 0.001, H_0 : \rho = 0$ )					
$\rho$	0			0.149		

Figure 3.5에서 보는 바와 같이 박탈지수를 공변량으로 고려하지 않았을 때는 소지역인 동별로 질병사망에 대한 SMR 추정치가 뚜렷한 차이를 가지고 있는 반면에, 공변량으로 고려하면 이러한 차이가 사라지는 것을 알 수 있다. 즉, 각 박탈지수가 심장질환 사망에 대한 SMR의 동별 차이를 잘 설명하고 있음을 의미한다.

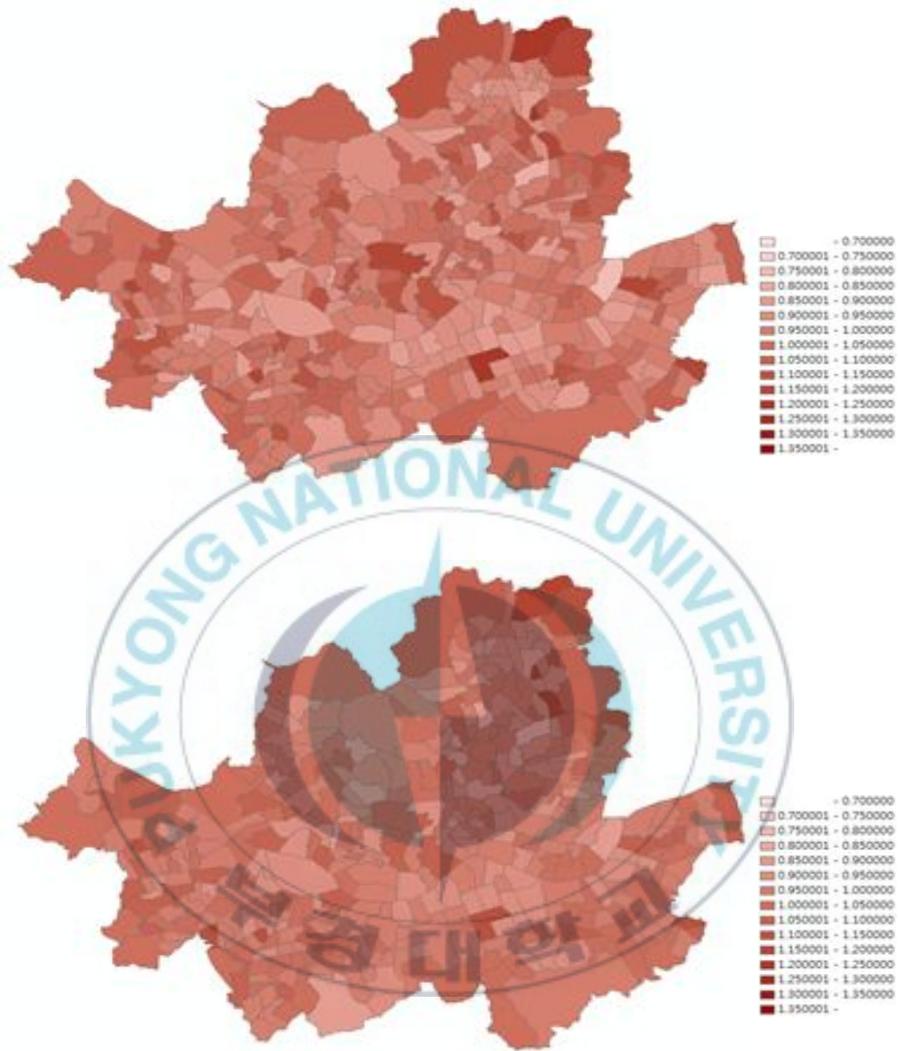


Figure 3.5 심장질환 사망에 대한 MRF모형 적합 후 SMR 추정치

(상 : 공변량 없음, 하 : 공변량 고려)

## 4. 결론 및 제언

본 연구에서는 서울의 행정 소지역 단위인 동별 주요 사망원인 중 질병, 사고, 암, 뇌혈관질환, 심장질환에 따른 사망자료를 활용하여 공간모형을 적합한 후 질병지도 결과를 제시하고자 하였다. 추정결과에서 보듯이 서울지역 동별 사망에 관한 자료는 동별 인접 지역 간 상관성이 존재한다는 것을 알 수 있다. 또한 서울지역 동별 박탈지수의 차이는 각 사망원인에 따라 유의한 영향을 미친다. 박탈지수를 고려하지 않았을 때는 서울 동별 지역에 따른 SMR의 차이가 뚜렷하게 나타나지만, 박탈지수를 고려하면 이러한 차이가 사라진다. 즉, 지역 간 박탈지수의 차이는 지역 간 건강수준의 차이에 뚜렷한 영향을 미치고 있음을 알 수 있다. 소지역별 사망률의 차이가 해당지역의 여러 요인에 의한 것으로 볼 수 있지만, 인접지역 요인에 따른 영향력도 의미 있다고 볼 수 있다.

본 연구의 결과로 서울지역의 질병, 사고, 암, 뇌혈관질환, 심장질환으로 인한 사망률이 높은 취약지역에 대한 관심과 예방대책이 마련될 수 있는 유용한 자료가 될 것으로 기대된다. 특히, 여러 사망원인에 대한 서울 외 다른 지역을 대상으로 한 다양한 소지역 연구들을 위한 연구방법에 본 연구에서 제시하는 통계적 방법론이 크게 기여할 것으로 기대된다.

## 참고문헌

- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004), *Hierarchical modelling and analysis for spatial data*, London: Chapman and Hall.
- Besag, J. E., York, J. and Molli, A. (1991), “Bayesian image restoration, with two applications in spatial statistics,” *Annals of Institute of Statistical Mathematics*, 43, 1–59.
- Clayton, d., Kaldor, J. (1987), “Empirical Bayes estimates of age-standardized relative risk for use in disease mapping,” *Biometrics*, 43, 671–681.
- Choi, M. H., Cheong, K. S. and Cho, B. M. (2011), “Deprivation and mortality at the town level in Busan: an ecological study,” *Journal of the Preventive Medicine and Public Health*, 44, 242–8.
- Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998), “Generalized linear models for small-area estimation,” *Journal of the American Statistical Association*, 93, 273–282.
- Kim, K. H., Noh, M. and Ha, I. D. (2011), “A study using HGLM on regional difference of the dead due to injuries,” *Journal of the Korean Data & Information Science Society*, 22, 137–148.

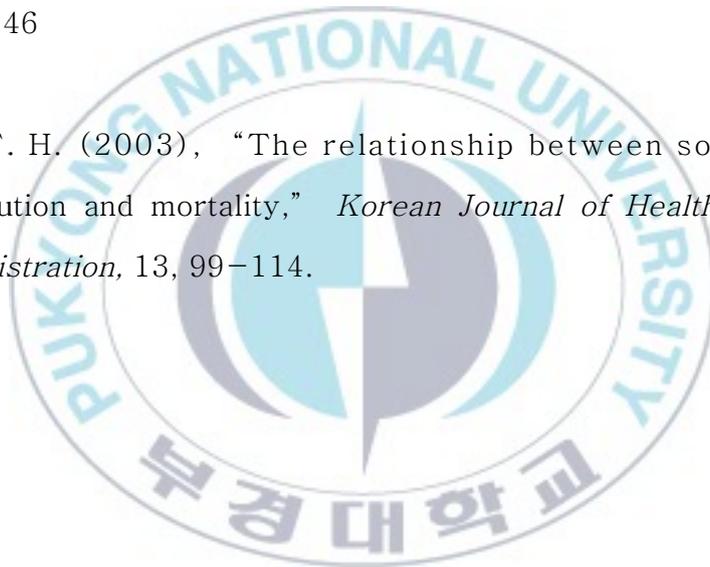
- Lee, Y. and Nelder, J. A. (1996), “Hierarchical generalized linear models(with discussion),” *Journal of the Royal Statistical Society B*, 58, 619–678.
- Lee, Y. and Ha, I. D. (2010), “Orthodox BLUP versus h-likelihood methods for inferences about random effects in tweedie mixed models,” *Statistics and Computing*, 20, 295–303.
- Lee, Y. and Nelder, J. A. (2001), “Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions,” *Biometrika*, 88, 987–1006.
- Lee, Y., Nelder, J. A and Pawitan (2006), *Generalized linear models with random effects: unified analysis via h-likelihood*, London: Chapman and Hall.
- Noh M. and Lee, Y. (2011), dhglm: Double hierarchical generalized linear models. R package version 1.0, URL <http://CRAN.R-project.org/package=dhglm>.
- Park, J. T. and Lee, S. E. (2001), “A comparative study of small area estimation methods,” *Journal of the Korea Data & Information Science Society*, 12, 47–55.

Rao, J. N. K. (2003), *Small area estimation*, New York: Wiley.

Rousset, F. (2014), spaMM: Mixed models, particularly spatial GLMMs. R package version 1.1, URL <http://CRAN.R-project.org/package=spaMM>.

Townsend, P. (1987), “Deprivation,” *Journal Social Policy*, 16, 125–146

Yoon, T. H. (2003), “The relationship between social class distribution and mortality,” *Korean Journal of Health Policy & Administration*, 13, 99–114.



## 부 록

1) seoul data file

: seoul\_data.csv

(1 가락1동, 2 가락2동, 3 가락본동, ... , 411 흑석동)

---

필드명	변수설명
region	서울 411개동 번호(1, 2, ... , 411)
score_ttl	박탈지수
obs_in	질병사망 관찰사망자수
exp_in	질병사망 기대사망자수
obs_out	사고사망 관찰사망자수
exp_out	사고사망 기대사망자수
obs_ca	암사망 관찰사망자수
exp_ca	암사망 기대사망자수
obs_brn	사고사망 관찰사망자수
exp_brn	사고사망 기대사망자수
obs_hrt	심장질환 사망 관찰사망자수
exp_hrt	심장질환 사망 기대사망자수

---

2) nbr

:  $411 \times 411$  neighborhood matrix

```
nbr[1,c(192,164,318,212,117,3,209)]<-1
```

```
nbr[2,c(3,116,337,19,284)]<-1
```

⋮

```
nbr[411,c(153,174,54)]<-1
```

# (1) 질병사망

```
seoul$obs_in<-round(seoul$obs_in)
```

```
res1_3<-HLfit(obs_in~score_ttl+(1|region)+offset(log(seoul$exp_in)),family=poisson(log),
```

```
HLmethod="HL(1,1)",rand.family=gaussian(),data=seoul)
```

```
res1_4<-corrHLfit(obs_in~score_ttl+adjacency(1|region)+offset(log(seoul$exp_in)),
```

```
data=seoul,family=poisson(log),
```

```
adjMatrix=nbr,lower=list(rho=-1),upper=list(rho=1))
```

# (2) 사고사망

```
seoul$obs_out<-round(seoul$obs_out)
```

```
res2_3<-HLfit(obs_out~score_ttl+(1|region)+offset(log(seoul$e
```

```

xp_out)),family=poisson(log),
HLmethod="HL(1,1)",rand.family=gaussian(),data=seoul

res2_4<-corrHLfit(obs_out~score_ttl+adjacency(1|region)+offs
et(log(seoul$exp_out)),
data=seoul,family=poisson(log),
adjMatrix=nbr,lower=list(rho=-1),upper=list(rho=1))

```

# (3) 암사망

```

seoul$obs_ca<-round(seoul$obs_ca)

res3_3<-HLfit(obs_ca~score_ttl+(1|region)+offset(log(seoul$e
xp_ca)),family=poisson(log),HLmethod="HL(1,1)",rand.family=g
aussian(),data=seoul)

res3_4<-corrHLfit(obs_ca~score_ttl+adjacency(1|region)+offse
t(log(seoul$exp_ca)),
data=seoul,family=poisson(log),
adjMatrix=nbr,lower=list(rho=-1),upper=list(rho=1))

```

# (4) 뇌혈관질환

```

seoul$obs_brn<-round(seoul$obs_brn)

```

```
res4_3<-HLfit(obs_brn~score_ttl+(1|region)+offset(log(seoul
$exp_brn)),family=poisson(log),
HLmethod="HL(1,1)",rand.family=gaussian(),data=seoul)
```

```
res4_4<-corrHLfit(obs_brn~score_ttl+adjacency(1|region)+offs
et(log(seoul$exp_brn)),
data=seoul,family=poisson(log),
adjMatrix=nbr,lower=list(rho=-1),upper=list(rho=1))
```

# (5) 심장질환

```
seoul$obs_hrt<-round(seoul$obs_hrt)
```

```
res5_3<-HLfit(obs_hrt~score_ttl+(1|region)+offset(log(seoul$e
xp_hrt)),family=poisson(log),
HLmethod="HL(1,1)",rand.family=gaussian(),data=seoul)
```

```
res5_4<-corrHLfit(obs_hrt~score_ttl+adjacency(1|region)+offs
et(log(seoul$exp_hrt)),
data=seoul,family=poisson(log),
adjMatrix=nbr,lower=list(rho=-1),upper=list(rho=1))
```