이 종 협 교수지도

석사학위청구논문

환경요인들이 뇌혈관 질환 사망자에 미치는 영향에 대한 시계열 분석

2009

성신여자대학교 대학원

통 계 학 과

박 희 원

환경요인들이 뇌혈관 질환 사망자에 미치는 영향에 대한 시계열 분석

이 종 협 교수지도

이 논문을 석사학위논문으로 제출함

2008년 11월

성신여자대학교 대학원

통 계 학 과

박 희 원

인 준 서

박희원의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

20세기의 마지막 30년 동안 우리나라의 인구는 45% 증가 하였고, 50%이던 도시화율은 80%까지 진행되는 등 급격한 산업화로 인해 최근 환경오염의 피해가 급증하고 있다. 심각한 환경오염 문제 중 대기오염은 다른 환경오염과 달리 대기오염물질이 한번 대기 중에 배출되면 제거할 방법이 전혀 없기 때문에 인체와 환경에 큰 영향을 미치는 것으로 알려져 있어 최근 더욱 관심이 집중되고 있다.

따라서 본 논문에서는 환경요인과 한국인의 3대 주요 사인 중 하나인 뇌혈관 질환 사망자 사이의 연관성을 규명하고자 1995년부터 2004년까지의 사망자 자료를 대상으로 분해모형, 자기회귀오차를 가지는 회귀모형, 자기회귀모형, 자기회귀 시차분포 모형을 적합 시켜 분석하고 향후 추이를 예측한다. 또한 시계열적 접근이 아닌 음이항 회귀모형을 통하여 환경요인이 뇌혈관 질환 사망자에 미치는 영향을 규명한다.

분석결과 뇌혈관 질환 사망자에 대한 최적모형으로는 AIC와 SBC통계량을 기준으로 ADL모형이 선택되었으며 아황산가스농도가 뇌혈관 질환 사망자에 유의한 영향을 미치는 것으로 나타났다. 반면 예측오차에 근거한 RMSE와 MAPE를 기준으로 비교 시 자기회귀오차를 가지는 회귀모형이 가장 우수했으며, 서울시 월별 뇌혈관 질환 사망자 자료는 추세성분은 나타나지 않고 계절성분은 유의한 것으로 나타났다. 음이항 회귀모형의 적합결과 아황산가스농도, 오존농도, 이산화질소농도, 평균기온, 평균상대습도가 뇌혈관 질환 사망자에 영향을 미치고 있다.

목 차

제1장 서 론

1900년대 이르러 급격한 산업화로 인해 환경오염에 의한 피해가 속출하고 있으며 그 중 대기오염은 다른 환경오염문제와는 달리 대기오염물질이 한번 대기 중에 배출되면 제거할 방법이 전혀 없어 인체와 자연환경에 미치는 영향이 매우 큰 것으로 알려져 있다. 또한 대규모 인구집단이 피해에 노출된다는 특성 때문에 더욱 치명적인 것으로 인식되고 있으며, 과거 뮤즈계곡 사건(1930, 벨기에), 요코하마 천식 사건(1976, 일본), 런던스모그 사건(1952, 영국), 보필 사건(1984, 인도) 등 대기오염으로 인한 대규모 피해의 발생으로 대기오염이 인체에 미치는 영향에 대한 관심이 고조되면서 이에 대한 연구가 국내·외에서 활발히 진행되고 있다.

1.1 대기오염물질과 일별 사망자 사이의 연관성 고찰

대기오염이 사망자수에 미치는 영향을 분석한 대부분의 연구에서는 일반화 가법모형(GAM: Generalized Additive Model)을 이용하여 분석을 시도하였다. 대기오염물질이 사망자수에 미치는 영향에 대해 국내에서 진행 된 연구는 다음과 같은 것들이 있다.

권호장과 조수현(1999)은 1991년 1월 1일부터 1995년 12월 31일까지의 서울시 사망자수와 기상변수(기온, 상대습도), 대기오염물질[SO_2(아황산가스), TSP(분진농도), NO_2(이산화질소), O_3(오존)]을 사용한 일반화 가법모형을 통하여 대기오염물질이 서울시 일별 사망자수에 미치는 영향에 대해 분석하였

다. 분석결과는 다음의 표 1.1과 같다.

표 1.1 분석결과(권호장, 조수현, 1999)

종속변수	영향을 미치는 변수
사망자	$O_{3,t-1}, SO_{2,t-2}, TSP_{t-2}, NO_{2,t-1}$
65세 미만 사망자	유의한 변수 없음
65세 이상 사망자	$O_{3,t-1}, SO_{2,t-1}, TSP_{t-1}, NO_{2,t-1}$
호흡기질환 사망자	$O_{3,t-5}, SO_{2,t}, TSP_{t}, NO_{2,t-1}$
순환계질환 사망자	명확하게 기재하지 않음

※ 변수$_{t-i}$: i번째 시차변수

김윤신 등(1998)은 1991년 1월 1일부터 1996년 12월 31까지의 울산, 여천지역 일별사망자수와 기상자료(기온,상대습도,풍속), 대기오염물질[SO_2, TSP, NO_2, O_3, CO(일산화탄소)]을 사용한 일반화 가법모형을 통하여 대기오염물질이 일별사망자수 증가에 어떠한 영향을 미치는가를 분석하였다. 분석결과는 다음의 표 1.2와 같다.

표1.2 분석결과(김윤신 등, 1998)

종속변수	영향을 미치는 변수
울산시 사망자	$O_{3,t-3}, SO_{2,t-2}, NO_{2,t-1}, CO_{t-3}$
여천시 사망자	$O_{3,t-5}, NO_{2,t-3}$

※ 변수$_{t-i}$: i번째 시차변수

양희은(2004)은 1991년 1월 1일부터 1999년 12월 31일까지 서울시 사망자수와 기상요인(기온, 상대습도, 풍속), 대기오염물질[SO_2, TSP, NO_2, O_3,

CO, PM_{10}(미세먼지)]을 사용한 일반화가법모형을 통하여 대기오염물질이 서울시 사망자수에 미치는 영향을 분석하였다. 분석결과는 다음의 표 1.3과 같다.

표 1.3 분석결과(양희은, 2004)

종속변수	영향을 미치는 변수
65세 이상 사망자	$PM_{10,t-2}, O_{3,t-3}, SO_{2,t-1}, NO_{2,t}, CO_{t-1}$
호흡기질환 사망자	$PM_{10,t}, O_{3,t-4}, SO_{2,t-3}, NO_{2,t}, CO_t$
순환기질환 사망자	$PM_{10,t-2}, O_{3,t-2}, SO_{2,t-2}, NO_{2,t-2}, CO_t$

※ 변수$_{t-i}$: i번째 시차변수

이승욱(2004)은 1991년 1월 1일부터 2002년 12월 31일까지의 부산지역의 일별사망자수와 기상자료(평균기온, 상대습도, 평균풍속), 대기오염물질[SO_2, NO_2, O_3, CO, PM_{10}]을 사용한 일반화가법모형을 통하여 일별 대기오염물질의 농도변화와 사망자수 증감의 관련성을 파악하였다. 분석결과는 다음의 표 1.4와 같다.

표 1.4 분석결과(이승욱, 2004)

종속변수	영향을 미치는 변수
사망자	$PM_{10,t-2}, SO_{2,t}, NO_{2,t-5}$
65세 미만 사망자	$NO_{2,t}$
65세 이상 사망자	$PM_{10,t}, SO_{2,t}, NO_{2,t}, CO_t, O_{3,t}$
호흡기질환 사망자	$SO_{2,t}, NO_{2,t-2}$
순환계질환 사망자	$SO_{2,t-5}, O_{3,t}$

※ 변수$_{t-i}$: i번째 시차변수

위 기존연구들은 모두 일별 사망자수가 포아송분포를 따른다고 가정하고 일반화 가법모형을 이용하여 평활함수(lowess)로 장기변동과 추세변동, 기상요인의 교란효과를 통제한 상태의 기본모형을 설정하였다. 이 기본모형에 대기오염물질을 하나씩 포함하여 단일오염물질의 효과를 보고자 하였다. 그러나 '장기변동 및 추세변동과 기상요인의 효과를 통제 한다'는 기존연구의 가정은 실제상황에서는 불가능 하며, 일별사망자수가 포아송 분포를 따른다는 가정을 뒷받침해줄 수 있는 통계적 근거를 제시하지 않았다.

따라서 먼저 기존분석에서 사용된 1995년 1월 1일부터 2004년 12월 31일까지의 일별 호흡기 질환 사망자 자료에 대해 포아송분포의 적합성검정을 실시하였다.

표 1.5 서울시 일별 호흡기 질환 사망자 적합도 검정

Y_t	확률	관측빈도	기대빈도
0	0.0042	20	15
1	0.0230	119	84
2	0.0630	264	230
3	0.1149	426	420
4	0.1571	595	574
5	0.1719	582	628
6	0.1567	525	572
7	0.1224	406	447
8	0.0837	280	306
9	0.0509	174	186
10	0.0278	122	102
11	0.0138	60	50
12	0.0063	27	23
13	0.0027	20	10
14이상	0.0016	33	6
Total		3653	

표 1.5에 주어진 카이제곱 검정결과 χ^2=178.9723으로 $p \leq 0.01$이므로 일별

호흡기 질환 사망자가 포아송분포를 따른다고 할 수 없다. 따라서 '일별 사망자수가 포아송 분포를 따른다'는 가정하에서의 일반화 가법 포아송 모형의 적합결과 역시 신뢰할 수 없으며 앞에서 언급했던 바와 같이 기존연구는 기상요인, 추세변동, 계절변동이 통제된다는 가정 하에서 하나의 대기오염물질의 독립적인 영향을 측정한 것에 불과하다.

1.2 뇌혈관 질환 사망자와 환경요인과의 연관성 연구방법

본 논문에서는 환경요인들이 뇌혈관 질환 사망자에 미치는 영향을 규명하고자 한다. 뇌혈관 질환으로 인한 사망은 한국인의 3대 주요사인 중 하나로 최근 이에 대한 관심이 높아지고 있어 본 논문에서는 뇌혈관 질환 사망자 자료를 대상으로 분석을 실시하였다. 또한 환경요인으로는 대기오염물질 [SO_2, PM_{10}, NO_2, O_3]과 기상요인(평균기온, 평균상대습도)을 고려하였다.

단일오염물질의 독립적인 효과를 측정하는 것이 아닌 기상요인과 추세변동 및 계절변동을 모두 고려한 현실적인 상황에서 대기오염물질이 사망자에 주는 영향을 파악하고자 다양한 시계열모형과 음이항 회귀모형을 통해 연관성을 규명하고, 뇌혈관 질환 사망자의 향후 추이를 살펴보고자 한다. 또한 기존연구는 일별사망자 자료를 사용하여 분석을 실시하였지만 본 논문에서는 통계청을 통해 일반인들이 쉽게 접근할 수 있는 서울시 월별자료를 사용하여 분석을 실시하였다.

본 논문에 사용된 시계열 모형은 추세성분, 계절성분, 순환성분, 불규칙 성분을 가지는 분해모형, 자기회귀오차를 가지는 회귀모형(autoregressive regression model with errors), 자기회귀모형(autoregressive model)과 자

기회귀 시차분포모형(autoregressive distributed lag model: ADL)이다.

본 논문의 구성은 다음과 같다. 제2장에서는 분해모형, 자기회귀오차를 가지는 회귀모형, 자기회귀모형, 자기회귀 시차분포모형을 이용하여 환경적 요인들이 서울시 월별 뇌혈관 질환 전체 사망자와 65세 이상 사망자에 미치는 영향을 분석한다. 제3장에서는 시계열적 접근이 아닌 음이항 회귀모형을 통하여 환경적 요인들의 영향을 분석한다. 제4장에서는 서울시 월별 전체 뇌혈관 질환 사망자에 대한 자기회귀오차를 가지는 회귀모형, 자기회귀모형, 자기회귀 시차분포모형을 이용하여 예측을 실시 한 후 예측력을 비교하고, 본 연구의 결론을 맺는다.

제2장 환경요인들이 서울시 월별 뇌혈관 질환 사망자에 미치는 영향 분석

뇌혈관 질환은 표 2.1에 나타난 바와 같이 암, 심장병과 함께 한국인의 3대 주요사망원인 질환 중 하나이며, 짧은 시간 내에 급격하게 질병의 경과가 진행하여 사망의 위험이 높은 질병으로 알려져 있다. 따라서 본 논문에서 사용한 사망자 자료는 1995년 1월부터 2004년 12월까지의 한국인의 3대 주요사망원인 질환인 뇌혈관 질환의 사망자를 대상으로 하였으며 120개의 월 사망자수를 대상으로 분석을 시도하였다.

표 2.1 한국인의 주요사망원인

	2000년	2001년	2002년	2003년	2004년
1위	신생물	신생물	신생물	신생물	신생물
2위	순환기계통의 질환	순환기계통의 질환	순환기계통의 질환	순환기계통의 질환	순환기계통의 질환
3위	뇌혈관 질환	뇌혈관 질환	뇌혈관 질환	뇌혈관 질환	뇌혈관 질환
4위	달리 분류되지 않은 증상, 징후	질병이환 및 사망의 외인	질병이환 및 사망의 외인	질병이환 및 사망의 외인	질병이환 및 사망의 외인
5위	질병이환 및 사망의 외인	달리 분류되지 않은 증상, 징후	달리 분류되지 않은 증상, 징후	달리 분류되지 않은 증상, 징후	달리 분류되지 않은 증상, 징후

뇌혈관 질환의 종류에는 뇌 동맥류, 뇌동맥 기형, 고혈압성 뇌출혈, 뇌경색, 모

야모야 병 등이 있는데 본 논문에서 사용한 사인은 표 2.2와 같으며 설명변수로 사용되는 환경요인은 표 2.3과 같다

표 2.2 분석에 사용된 사인

질환명	소분류
뇌혈관 질환	-순환기계통의 질환(I00-I99) -대뇌혈관 질환(I60-I69) -거미막밑 출혈(I60) -뇌내출혈(I61) -기타 비외상성 머리내 출혈(I62) -뇌경색증(I63) -출혈 또는 경색증으로 명시되지 않은 뇌중풍(I64) -대뇌경색증을 유발하지 않은 뇌전동맥의 폐색 및 협착(I65) -대뇌경색을 유발하지 않은 대뇌동맥의 폐색 협착(I66) -기타 뇌혈관 질환(I67) -달리 분류된 질환에서의 뇌혈관 장애(I68) -뇌혈관 질환의 휴유증(I69)

※() : 한국 표준 질병 사인분류(Korea Classification of Diseases; KCD)

표 2.3 분석에 사용된 환경요인

변수명		표기법	변수설명
대기오염물질	아황산가스(SO_2)	X1	통계청에서 제공하는 대기오염도 자동측정자료로 단위는 ppm이다.
	오존(O_3)	X2	통계청에서 제공하는 대기오염도 자동측정자료로 단위는 ppm이다.
	이산화질소(NO_2)	X3	통계청에서 제공하는 대기오염도 자동측정자료로 단위는 ppm이다.
	미세먼지(PM_{10})	X4	통계청에서 제공하는 대기오염도 자동측정자료로 단위는 $\mu g/m^3$이다.
기상요인	평균기온	X5	통계청에서 제공하는 기후자료로 월평균수치이며 단위는 ^oC이다.
	평균상대습도	X6	통계청에서 제공하는 기후자료로 월평균수치이며 단위는 %이다.

위 대기오염물질 자료는 1995년 1월부터 2004년 12월까지 서울시의 20여개 측정소에서 관측된 값으로 시간별 데이터(data)를 수집해, 각 시간대 별 20여개 측정소의 평균값을 산출한 후 하루 24개의 평균값을 이용하여 일 평균값을 구하고, 이를 가지고 월 평균을 구한 값으로 통계청에서 제공하는 자료이다. 또한 기상요인은 서울시 종로구에 있는 기상측정소에서 3시간 마다 측정된 값을 수집해 하루 8개의 값으로 일 평균값을 구하고, 이를 이용해 월 평균값을 구한 것으로 이 역시 통계청에서 제공하는 자료이다. 위의 X1부터 X6까지의 설명변수 역시 120개의 월 자료를 대상으로 분석을 실시하였다.

따라서 위 표 2.2에 나타나 있는 사인의 월별 사망자 자료를 대상으로 분해모형, 자기회귀오차를 가지는 회귀모형, 자기회귀모형을 적합하고, 자기회귀모형에 표 2.3의 환경요인을 고려한 ADL모형과 음이항 회귀모형을 적합 한 후 시계열 모형을 통해 사망추이를 예측한다.

2.1 시계열모형

본 논문에서 환경요인과 뇌혈관 질환 사망자의 연관성 규명에 사용된 시계열 모형에 대해 간략히 설명하면 다음과 같다.

2.1.1 분해모형

분해기법은 시계열 자료 Y_t를 구성하고 있는 성분들이 결정적으로 서로 독립이라 가정하고 Y_t를 추세성분(T_t), 계절성분(S_t), 순환성분(C_t), 불규칙 성분(I_t)으로 분해한 후 이를 이용하여 미래를 예측하는 방법이다. 위 시계열 성분들이 시계열에 가법적 또는 승법적으로 표현되었는가에 따라 분해모형을 다음과 같은 두 가지 형태로 구분한다.

가법모형(Addictive Model) : $Y_t = T_t + S_t + C_t + I_t,$

승법모형(Multiplicative Model) : $Y_t = T_t \cdot S_t \cdot C_t \cdot I_t.$

2.1.2 자기회귀모형

자기회귀모형은 시계열 자체에 대한 회귀형태를 취하는 모형으로 일반 p차 AR모형을 따르는 Y_t'는 다음과 같이 나타낸다.

$$Y_t' = \phi_1 Y'_{t-1} + \phi_2 Y'_{t-2} + \cdots + \phi_p Y'_{t-p} + a_t,$$

여기서 $Y_t' = Y_t - \mu$, $a_t = WN(0,\sigma^2)$. 현시점 t에서의 시계열 Y_t'는 p개의 과거 값들의 가중 합과 이들로 설명되지 않는 부분인 오차항 a_t의 선형결합으로 표현된다.

2.1.3 자기회귀 시차분포모형(ADL모형)

Y_t의 시차변수들 이외에 다른 설명변수들의 시차변수들이 자기회귀모형에 포함될 때 이를 자기회귀 시차 분포모형(ADL)이라한다. k개의 추가적인 설명변수들 중 X_1은 q_1개의 시차변수, X_2는 q_2개의 시차변수,…, X_k는 q_k개의 시차변수가 포함되어 있다고 가정하면 $ADL(p,q_1,\cdots,q_k)$모형은 다음과 같이 표현된다.

$$\begin{aligned} Y_t = \alpha &+ \delta_1 Y_{t-1} + \cdots + \delta_p Y_{t-p} \\ &+ \beta_{11} X_{1,t-1} + \ldots + \beta_{1q_1} X_{1,t-q_1} \\ &+ \qquad\qquad \cdots \\ &+ \beta_{k1} X_{k,t-1} + \ldots + \beta_{kq_k} X_{k,t-q_k} + e_t. \end{aligned}$$

여기서 α, δ, β는 미지의 계수이고 e_t는 오차 항으로 $E(e_t | Y_{t-1}, Y_{t-2} \cdots, X_{1,t-1}, \cdots X_{k,t-q_k}) = 0$이다. 확률변수 $(Y_t, X_{1,t}, X_{2,t}, \cdots, X_{k,t})$는 정상성(stationary)분포를 가지고 있고, $(Y_t, X_{1t}, X_{2t}, \cdots X_{kt})$와 $(Y_{t-j}, X_{1,t-j}, \cdots, X_{k,t-j}, \cdots)$는 j가 커짐에 따라 상관관계가 소멸된다. 또한 $X_{1,t}, \cdots, X_{k,t}$와 Y_t는 0이 아닌 유한한 값을 가지고 있으며 완전한 다중공선

성이 없다고 가정한다. 일반적으로 이 모형에는 하나 이상의 종속변수의 시차변수가 포함된다(Stock and Watson, 2002).

2.1.4 모형선택기준

주어진 시계열 자료에서 여러 가지 후보 모형이 존재할 경우 일반적으로 사용되는 모형선택기준은 잔차 $\hat{a_t}$에 근거한 Akaike(1973, 1974)의 AIC(Akaike's Information Criterion)와 Schwartz(1978)의 SBC(Schwartz's Baysian Criterion) 등이 주로 사용되며 모형 선정 시 AIC와 SBC값을 최소로 하는 모형을 선택한다(Wei, 2006).

2.2 서울시 월별 뇌혈관 질환 사망자 모형

2.1절에서 언급한 다양한 시계열모형을 이용하여 월별 뇌혈관 질환 사망자 자료를 분석해 보겠다.

2.2.1 분해모형

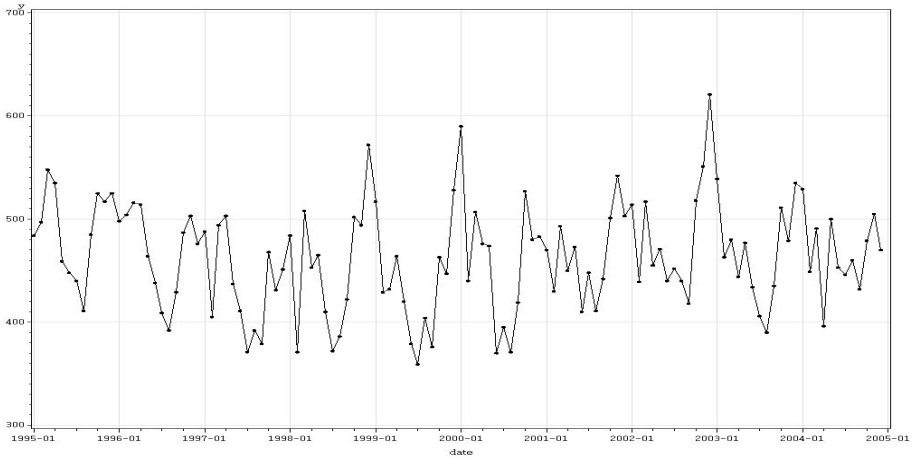


그림 2.1 서울시 월별 뇌혈관 질환 사망자

그림 2.1을 보면 서울시 월별 뇌혈관 질환 사망자는 계절변동이 선형추세에 의존하지 않고 규칙적이며 진폭이 일정한 형태를 띠고 있어 가법모형이 적절하다고 생각된다. 즉, $Y_t = T_t + S_t + I_t$를 따른다고 하자. 일반적으로 순환성분 C_t는 고려하지 않으므로 모형에서 제외하였다(이종협, 2007). 추세성분 t와 계절변동을 나타내는 ID_{t1}부터 ID_{t11}까지의 지시함수 ($t = i$월 이면 $ID_{ti} = 1$이고 그렇지않다면 $ID_{ti} = 0$)들을 포함하는 다음의 분해모형을 고려하자. 추정결과는 표 2.4와 같다.

$$Y_t = \beta_0 + \beta_1 t + \sum_{i=1}^{11} \gamma_i ID_{tj} + e_t.$$

표 2.4 분해모형 추정결과

Ordinary Least Squares Estimates

SSE	127003.003	DFE	107
MSE	1187	Root MSE	34.45205
SBC	1238.51955	AIC	1202.28216
MAE	25.0511111	AICC	1205.71612
MAPE	5.45468382	Regress R-Square	0.5962
Durbin-Watson	1.2337	Total R-Square	0.5962

Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	1.2337	<.0001	1.0000

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	513.3139	12.4484	41.24	<.0001
t	1	0.0468	0.0912	0.51	0.6094
jan	1	-4.5856	15.4401	-0.30	0.7670
feb	1	-73.2324	15.4344	-4.74	<.0001
mar	1	-17.3792	15.4293	-1.13	0.2625
apr	1	-47.0259	15.4247	-3.05	0.0029
may	1	-52.0727	15.4207	-3.38	0.0010
jun	1	-96.8194	15.4172	-6.28	<.0001
jul	1	-106.3662	15.4142	-6.90	<.0001
aug	1	-110.5130	15.4117	-7.17	<.0001
sep	1	-92.5597	15.4099	-6.01	<.0001
oct	1	-18.2065	15.4085	-1.18	0.2400
nov	1	-21.4532	15.4077	-1.39	0.1667

표 2.4의 분해모형 추정결과를 살펴보면 추세성분 t가 유의하지 않은 것으로 나타나 추세성분을 모형에서 고려하지 않았다. 추세성분 t를 고려하지 않은 최종 분해모형 식(2.1)의 최소제곱 추정결과가 표 2.5에 나타나 있다.

$$Y_t = \beta_0 + \sum_{i=1}^{11} \gamma_i ID_{tj} + e_t \tag{2.1}$$

표 2.5 최종 분해모형 추정결과

Ordinary Least Squares Estimates

SSE	127314.7	DFE	108
MSE	1179	Root MSE	34.33424
SBC	1234.02621	AIC	1200.57631
MAE	25.1766667	AICC	1203.4922
MAPE	5.48443313	Regress R-Square	0.5952
Durbin-Watson	1.2305	Total R-Square	0.5952

Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	1.2305	<.0001	1.0000

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	516.4000	10.8574	47.56	<.0001
jan	1	-5.1000	15.3547	-0.33	0.7404
feb	1	-73.7000	15.3547	-4.80	<.0001
mar	1	-17.8000	15.3547	-1.16	0.2489
apr	1	-47.4000	15.3547	-3.09	0.0026
may	1	-52.4000	15.3547	-3.41	0.0009
jun	1	-97.1000	15.3547	-6.32	<.0001
jul	1	-106.6000	15.3547	-6.94	<.0001
aug	1	-110.7000	15.3547	-7.21	<.0001
sep	1	-92.7000	15.3547	-6.04	<.0001
oct	1	-18.3000	15.3547	-1.19	0.2359
nov	1	-21.5000	15.3547	-1.40	0.1643

표 2.5의 최종 분해모형 추정결과를 살펴보면 오차들의 1차 자기상관 존재유무를 판단할 수 있는 DW(Durbin-Waston)통계량 값이 1.2305로 $p<0.0001$이므로 자기상관이 존재하는 것을 알 수 있다. 잔차들에 시도표(그림 2.2) 역시 자기상관의 경향을 보이고 있으므로 오차항 간의 자기상관을 고려한 자기회귀오차를 가지는 회귀모형을 적합 시키는 것이 올바른 것으로 보인다.

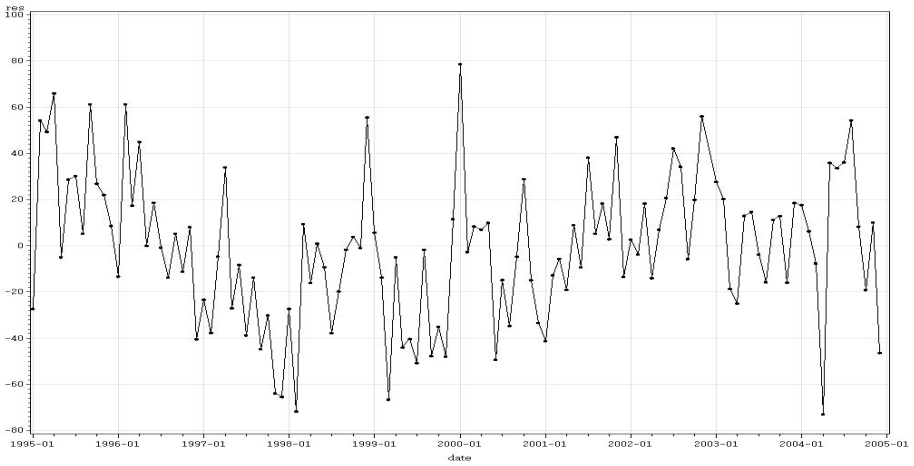


그림2.2 모형(2.1) 적합 후 잔차의 시도표

이 자료의 계절주기가 12이므로 이를 포함하는 충분한 시차를 고려한 자기회귀오차를 가지는 회귀모형을 적합 시킨 결과가 표 2.6에 주어져 있다. 이 과정에서 'BACKSTEP'이라는 SAS 명령어를 사용하여 오차의 유의하지 않은 차수를 제거하였으며, 최종적으로 차수 1, 2만이 유의한 것으로 판정되었다. 적합 후 잔차의 자기상관 존재여부를 판단하기 위한 DW검정의 p-값은 0.3935로써 더 이상의 자기상관이 존재하지 않는다. 자기회귀오차를 가지는 회귀모형은 식 (2.2)와 같다.

$$\begin{aligned}\hat{Y}_t &= 516.7897 - 6.6531ID_{t,jan} - 75.4265ID_{t,feb} - 18.8328ID_{t,mar} - 48.2672ID_{t,apr} \\ &\quad - 53.0755ID_{t,may} - 97.6878ID_{t,jun} - 107.1341ID_{t,jul} - 111.2001ID_{t,aug} \\ &\quad - 93.2379ID_{t,sep} - 18.7608ID_{t,oct} - 22.3406ID_{t,nov} + \hat{e}_t, \\ \hat{e}_t &= 0.2944\hat{e}_{t-1} + 0.2115\hat{e}_{t-2}. \end{aligned} \tag{2.2}$$

표 2.6 자기회귀오차를 가지는 회귀모형

Maximum Likelihood Estimates

SSE	104454.579	DFE	106
MSE	985.42055	Root MSE	31.39141
SBC	1220.09365	AIC	1181.06877
MAE	23.3836625	AICC	1185.06877
MAPE	5.05665447	Regress R-Square	0.5810
Durbin-Watson	1.9319	Total R-Square	0.6679

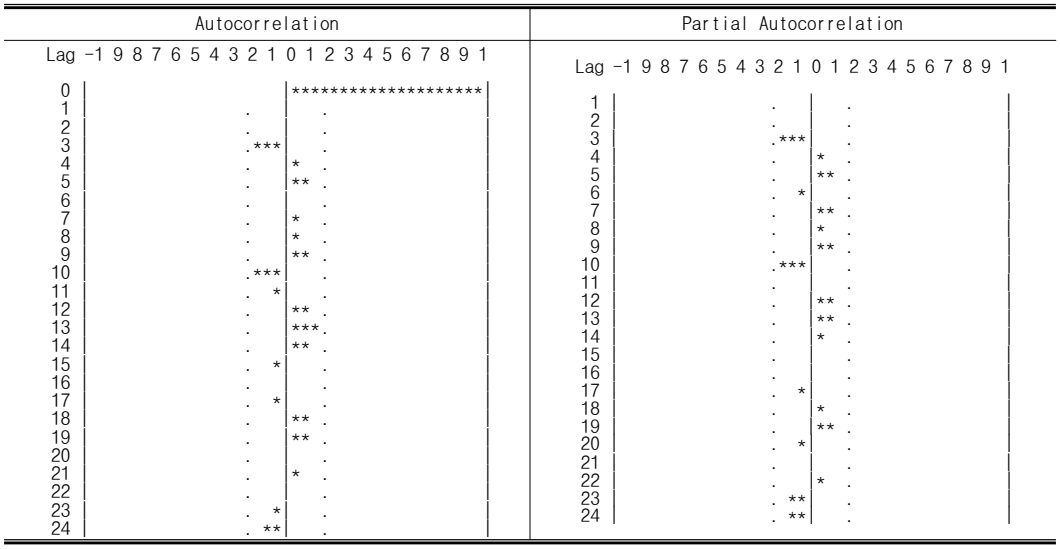
Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	1.9319	0.3935	0.6065

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	516.7897	10.8882	47.46	<.0001
jan	1	-6.6531	12.4567	-0.53	0.5944
feb	1	-75.4265	12.9088	-5.84	<.0001
mar	1	-18.8328	14.0858	-1.34	0.1841
apr	1	-48.2672	14.4514	-3.34	0.0012
may	1	-53.0755	14.7206	-3.61	0.0005
jun	1	-97.6878	14.7689	-6.61	<.0001
jul	1	-107.1341	14.6949	-7.29	<.0001
aug	1	-111.2001	14.3893	-7.73	<.0001
sep	1	-93.2379	13.9899	-6.66	<.0001
oct	1	-18.7608	12.7072	-1.48	0.1428
nov	1	-22.3403	12.2414	-1.82	0.0708
AR1	1	-0.2944	0.0960	-3.07	0.0027
AR2	1	-0.2115	0.0962	-2.20	0.0301

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.95	6	0.6838	0.022	0.008	-0.127	0.035	0.115	-0.014
12	10.41	12	0.5803	0.070	0.025	0.111	-0.146	-0.034	0.091
18	15.52	18	0.6261	0.133	0.083	-0.056	-0.019	-0.032	0.089
24	19.90	24	0.7024	0.098	-0.021	0.037	0.021	-0.075	-0.109



이 모형에 의해 추정된 추세·계절성분은 표 2.7에 주어져 있고, 불규칙성분은 원계열 Y_t에서 추세·계절성분의 차로 얻을 수 있다.

표 2.7 각 성분의 추정결과

날짜	원계열	추세·계절성분	불규칙성분
1995-01-01	484	510.14	-26.14
1995-02-01	497	431.61	65.40
1995-03-01	548	508.81	39.19
1995-04-01	535	495.02	39.98
1995-05-01	459	493.87	-34.87
(중략)			
2004-08-01	460	423.46	36.54
2004-09-01	432	447.26	-15.26
2004-10-01	479	512.02	-33.02
2004-11-01	505	490.63	14.37
2004-12-01	470	515.87	-45.87

2.2.2 자기회귀 시차분포(ADL)모형

모형의 설명력 및 예측력을 높이기 위해 Y_t의 시차변수들 뿐 아니라 다른 설명변수들의 시차변수들을 추가적으로 고려한 ADL모형에 적합하여 분석을 시도하였다.

2.2.2.1 자기회귀모형

표2.8 서울시 월별 뇌혈관 질환 사망자 SACF와 SPACF

Name of Variable = y

Mean of Working Series 462.7917
Standard Deviation 51.19341
Number of Observations 120

0 Autocorrelation

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	2620.765	1.00000	|********************|	0
1	1248.774	0.47649	|********** .	0.091287
2	721.217	0.27519	|****** .	0.110079
3	-144.850	-.05527	. *| .	0.115670
4	-621.330	-.23708	*****| .	0.115890
5	-495.685	-.18914	.****| .	0.119864
6	-676.713	-.25821	*****| .	0.122325
7	-432.573	-.16506	. ***| .	0.126786
8	-554.697	-.21165	.****| .	0.128564
9	-18.919575	-.00722	. | .	0.131436
10	374.257	0.14280	. |*** .	0.131439
11	862.812	0.32922	. |******* .	0.132726
12	1655.115	0.63154	. |************ .	0.139365
13	1018.929	0.38879	. |******** .	0.161462
14	508.971	0.19421	. |**** . .	0.169084
15	-262.821	-.10028	. **| . .	0.170932
16	-765.937	-.29226	.******| . .	0.171422
17	-647.579	-.24710	. *****| . .	0.175525
18	-616.245	-.23514	. *****| . .	0.178400
19	-360.426	-.13753	. ***| . .	0.180965
20	-482.844	-.18424	. ****| . .	0.181833
21	-66.452734	-.02536	. *| . .	0.183382
22	351.142	0.13398	. |*** . .	0.183412
23	579.306	0.22104	. |**** . .	0.184226
24	1228.526	0.46877	. |********* .	0.186423

Partial Autocorrelation

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.47649	. |********** .
2	0.06229	. |* . .
3	-0.27003	*****| . .
4	-0.18651	****| . .
5	0.08741	. |** . .
6	-0.16567	.***| . .
7	-0.06635	. *| . .
8	-0.15484	.***| . .
9	0.17246	. |***. .
10	0.13088	. |***. .
11	0.18025	. |**** .
12	0.48330	. |********** .
13	-0.08163	. **| . .
14	-0.20132	****| . .
15	-0.01138	. | . .
16	-0.06546	. *| . .
17	0.01389	. | . .
18	0.02707	. |* . .
19	-0.00291	. | . .
20	-0.10752	. **| . .
21	-0.01235	. | . .
22	0.15941	. |***. .
23	-0.10875	. **| . .
24	0.04948	. |* . .

표 2.8에 있는 Y_t의 SACF와 SPACF에 대한 유의성 검정과 패턴을 살펴보면 SPACF가 시차 1과 12에서 유의하며 SACF는 12의 배수에 해당하는 시차를

따라 감소하고 시차 1, 2, 3,…에 대해서도 감소하는 패턴을 보이므로 모형 $(1-\phi_1 B)(1-\Phi_1 B^{12})Y_t = \theta_0 + a_t$이 가능한 반면, SACF는 감소하는 패턴이고 SPACF는 시차 1, 2, 3, 12에서 유의한 것으로 간주하여 모형 $(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_{12} B^{12})Y_t = \theta_0 + a_t$이 적합할 수도 있다고 판단되어 두 모형을 적합 시켰다. 그 결과 AIC와 SBC통계량을 기준으로 비교하였을 때 각각 두 모형의 AIC통계량이 1208.662, 1203.123이고, SBC통계량이 1217.025, 1217.060로 큰 차이는 없었으나 해석의 용이성 측면에서 모형 $(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_{12} B^{12})Y_t = \theta_0 + a_t$을 고려한다.

모형$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_{12} B^{12})Y_t = \theta_0 + a_t$을 최우추정방법으로 추정한 결과가 표 2.9에 주어져 있다. 표 2.9를 살펴보면 모수추정치는 모두 유의한 것으로 나타났으며 위 2.2.1절의 자기회귀오차를 가지는 회귀모형과 비교 시 AIC통계량은 증가하였고 SBC통계량은 감소하였다. 한편 포토맨토우 검정(portmanteau test)결과 모형이 적합함을 보이고 있으며, 잔차의 SACF와 SPACF는 아무런 패턴을 보이지 않고 있으므로 식 (2.3)을 최종 자기회귀모형으로 선택하였다.

$$\underset{}{Y_t} = \underset{}{467.73920} + \underset{(0.07263)}{0.2584}\,Y_{t-1} + \underset{(0.07495)}{0.16231}\,Y_{t-2} - \underset{(0.06943)}{0.17984}\,Y_{t-3} + \underset{(0.06548)}{0.54641}\,Y_{t-12} \qquad (2.3)$$
$$(11.74559)$$

표 2.9 서울시 월별 뇌혈관 질환 사망자에 대한 자기회귀모형 추정결과

Maximum Likelihood Estimation

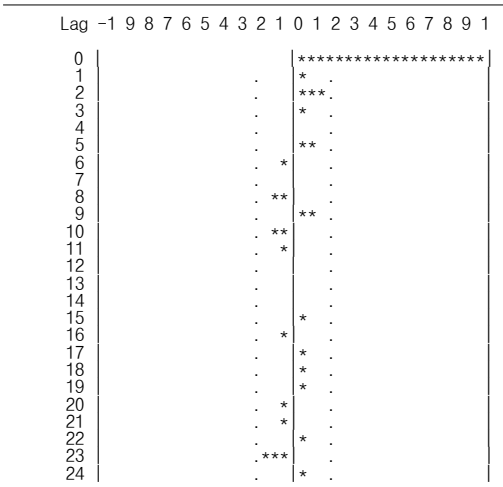
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	467.73920	11.74559	39.82	<.0001	0
AR1,1	0.25584	0.07263	3.52	0.0004	1
AR1,2	0.16231	0.07495	2.17	0.0304	2
AR1,3	-0.17984	0.06943	-2.59	0.0096	3
AR1,4	0.54641	0.06548	8.35	<.0001	12

Constant Estimate	100.6947
Variance Estimate	1212.395
Std Error Estimate	34.81947
AIC	1203.123
SBC	1217.06
Number of Residuals	120

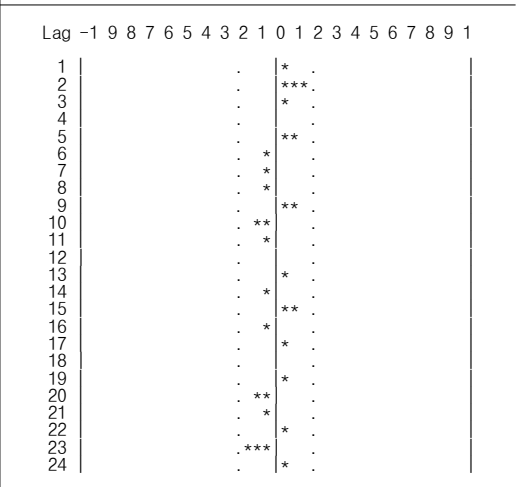
Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.08	2	0.0789	0.045	0.155	0.065	0.008	0.091	-0.045
12	8.85	8	0.3550	-0.013	-0.077	0.098	-0.106	-0.030	-0.024
18	9.87	14	0.7714	0.018	-0.003	0.054	-0.034	0.034	0.041
24	15.48	20	0.7483	0.034	-0.040	-0.047	0.059	-0.163	0.046

Autocorrelation



Partial Autocorrelation



2.2.2.2 자기회귀 시차분포(ADL)모형

모형의 개선을 위해 자기회귀모형에 환경요인을 추가적으로 고려한 ADL 모형에 적합하고자 한다. 모형에 고려할 환경요인 중 대기오염 물질은 5대 대기오염물질인 $SO_2, CO, O_3, NO_2, PM_{10}$중 통계청 제공하는 $SO_2(=X1)$, $O_3(=X2), NO_2(=X3), PM_{10}(=X4)$을 고려하였으며 기상요인으로는 평균기온($X5$) 및 평균상대습도($X6$)를 고려하였다. 대기오염물질의 농도 및 기상요인의 경우 과거 값이 당월 사망자수에 영향을 미칠 수 있으므로, 즉 지연효과가 있을 것으로 생각되어 과거 시차변수를 모형에 고려하고자 한다. 시차변수의 차수선택은 추후 분석결과에 매우 중요한 영향을 미친다. 시차변수의 차수를 선택 시 고려해야 할 사항은 효율성과 편의성인데, 차수의 길이가 길수록 편의는 줄어들지만 너무 많은 변수들을 모형에 포함하는 경우 효율성 측면에서는 좋지 못하다. 따라서 본 논문에서는 표 2.10의 서울시 월별 뇌혈관 질환 사망자와 대기오염물질 및 기상요인의 상관분석 결과 고차 시차변수들이 유의한 상관을 나타내고 있으므로 모형 선택의 효율성과 편의성을 고려하여 설명변수의 시차를 시차 5까지만 고려하고자 한다(Daniel and Dalls, 1985).

위와 같은 과정에 의해 모형에 포함하는 시차의 길이를 5로 선택한 후 그랜져 인과성검정(Granger, 1969)을 통해 대기오염물질과 기상요인들이 다른 변수들에 의한 예측력 이외에 추가적으로 유용한 예측력이 있는 지를 검정하였다.

표 2.10 상관분석 결과

		시차 1	시차 2	시차 3	시차 4	시차 5
		$X1_{t-i}(SO_2)$				
	Corr	0.2180	-0.0017	-0.2122	-0.3335	-0.3731
	p-value	(0.0177)	(0.9857)	(0.0222)	(0.0003)	(<.0001)
		$X2_{t-i}(O_3)$				
	Corr	-0.5940	-0.5179	-0.3807	-0.1059	0.1762
	p-value	(<.0001)	(<.0001)	(<.0001)	(0.2601)	(0.0607)
		$X3_{t-i}(NO_2)$				
	Corr	0.3611	0.0114	-0.2866	-0.3037	-0.2914
Y_t	p-value	(<.0001)	(0.9027)	(0.0018)	(0.0010)	(0.0017)
		$X4_{t-i}(PM_{10})$				
	Corr	0.1509	-0.1888	-0.3635	-0.3108	-0.2911
	p-value	(0.1028)	(0.0414)	(<.0001)	(0.0007)	(0.0017)
		$X5_{t-i}(Tem)$				
	Corr	-0.4670	-0.1939	0.1196	0.4045	0.5567
	p-value	(<.0001)	(0.0354)	(0.1991)	(<.0001)	(<.0001)
		$X6_{t-i}(Hum)$				
	Corr	-0.3806	-0.0263	0.3084	0.5205	0.4829
	p-value	(0.0001)	(0.7772)	(0.0007)	(<.0001)	(<.0001)

2.2.2.1절의 자기회귀모형에 $X1$부터 $X6$까지 변수의 시차변수들을 포함하여 그랜져 인과성 검정을 한 결과가 표 2.11에 주어져 있다. 검정결과 대기오염물질과 기상변수 중 변수 $X1(SO_2)$의 과거 값만이 사망자(Y_t)의 과거 값에 포착된 부분이외의 사망자 변화를 예측하는데 유용한 정보를 포함하고 있는 것으로 보인다. 따라서 2.2.2.1절의 자기회귀모형에 $X1(SO_2)$의 과거 값을 포함하는 ADL모형을 고려해 보도록 하겠다.

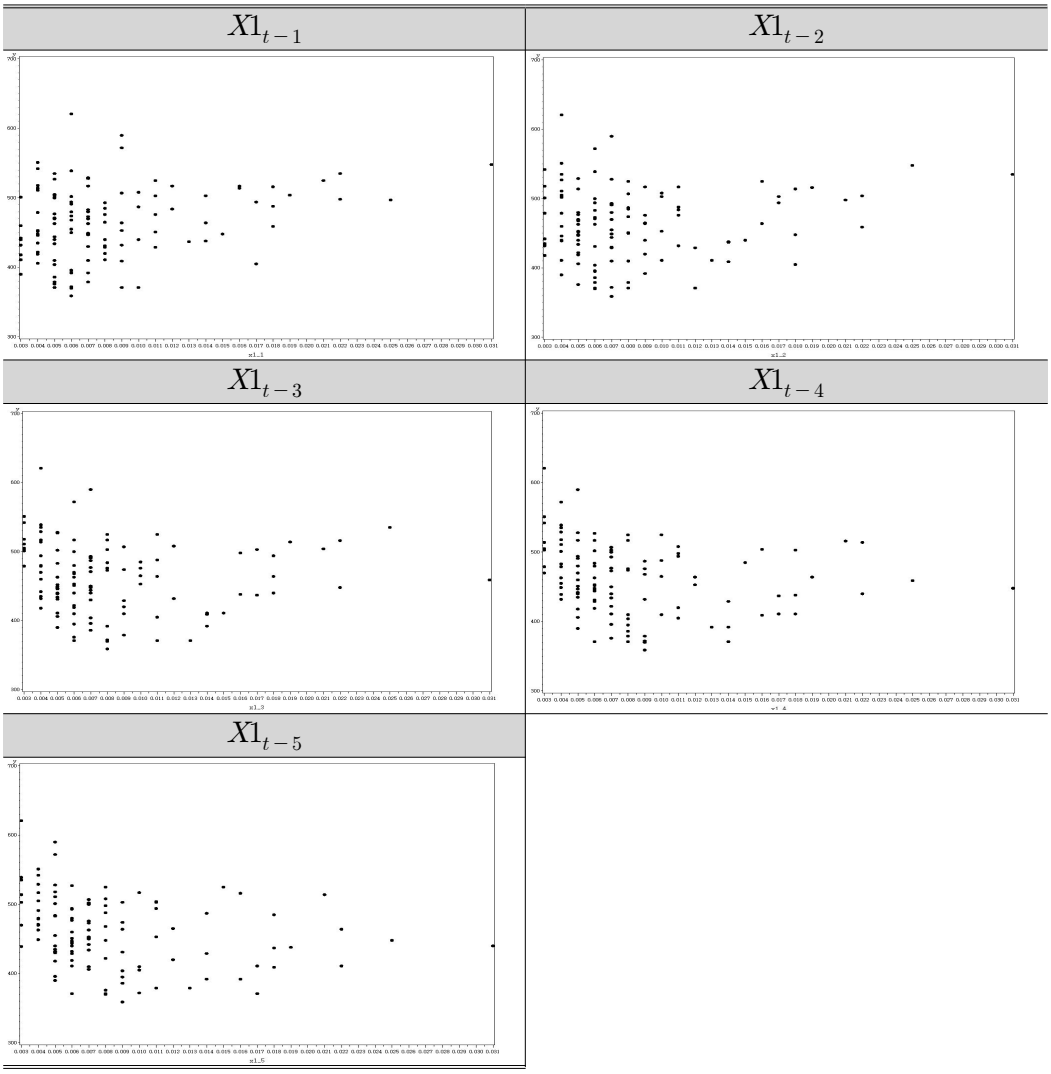


그림 2.3 서울시 월별 뇌혈관 질환 사망자와 $X1$의 상관 Plot

표 2.11 그랜져 인과성 검정(Granger casuality Test) 결과

변수	인과성검정(Granger casuality Test) 결과				
X1	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	3029.042849	2.56	0.0322
	Denominator	93	1181.548265		
X2	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	711.192521	0.54	0.7421
	Denominator	93	1306.163874		
X3	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	1093.688577	0.85	0.5174
	Denominator	93	1285.599570		
X4	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	1719.043943	1.39	0.2341
	Denominator	93	1234.087287		
X5	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	1297.142778	1.05	0.3923
	Denominator	98	1234.124428		
X6	Test granger				
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	5	1913.839292	1.59	0.1696
	Denominator	98	1202.660320		

자기회귀모형에 $X1(SO_2)$의 1-5시차의 과거 값을 포함한 ADL모형의 추정 결과가 표 2.12에 주어져 있다.

표 2.12 서울시 월별 뇌혈관 질환 사망자에 대한 ADL모형

Ordinary Least Squares Estimates

SSE	109883.989	DFE	93
MSE	1182	Root MSE	34.37366
SBC	1056.81113	AIC	1030.46384
MAE	26.7092914	AICC	1032.85514
MAPE	5.83364841	Regress R-Square	0.6084
		Total R-Square	0.6084

Miscellaneous Statistics

Statistic	Value	Prob	Label
Durbin's t	0.3977	0.3459	Pr > t

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	145.0255	55.6724	2.60	0.0107
y_1	1	0.2253	0.0850	2.65	0.0094
y_2	1	0.1615	0.0869	1.86	0.0663
y_3	1	-0.1721	0.0841	-2.05	0.0436
y_12	1	0.4973	0.0744	6.68	<.0001
x1_1	1	5033	2539	1.98	0.0504
x1_2	1	-8005	3355	-2.39	0.0191
x1_3	1	2952	3416	0.86	0.3898
x1_4	1	2322	3416	0.68	0.4984
x1_5	1	-4035	2358	-1.71	0.0904

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.20	6	0.6491	0.024	0.120	-0.008	-0.035	0.130	-0.106
12	12.91	12	0.3755	-0.100	-0.146	-0.000	-0.181	-0.063	-0.176
18	14.95	18	0.6655	0.004	0.004	0.115	0.002	0.099	0.043
24	20.16	24	0.6877	0.050	0.039	-0.123	0.060	-0.214	-0.017

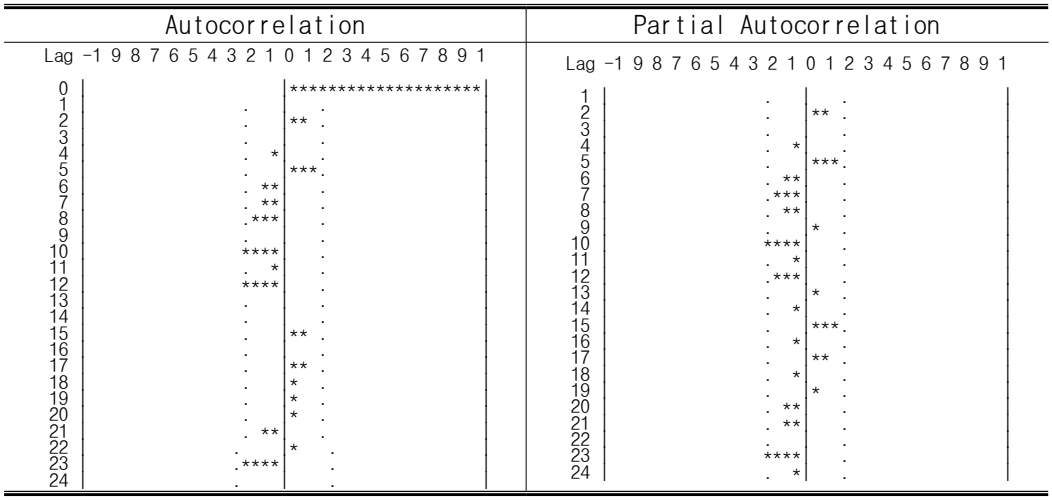


표 2.12의 ADL모형 추정결과 변수 $X1_{t-3}, X1_{t-4}, X1_{t-5}$가 유의하지 않은 것으로 나타나 이를 제거 후 모형을 적합 시킨 결과가 아래 표 2.13에 나타나 있다.

표 2.13의 ADL모형 추정결과를 살펴보면 앞의 자기회귀모형과 비교 시 모형 선택의 기준이 되는 통계량인 AIC가 1203.123에서 1055.183로, SBC가 1217.06에서 1073.827로 감소하였다. 또한 포토맨토우 검정과 잔차의 SACF와 SPACF 역시 모형이 적합함을 보이고 있으므로 식 (2.4)를 서울시 월별 뇌혈관 질환 사망자의 ADL 최종모형으로 선택하였다. 식 (2.4)를 보면 한달 전, 두달 전, 세달 전, 일년 전 뇌혈관 질환 사망자와 한달 전, 두달 전 아황산가스($X1$)농도가 서울시 월별 뇌혈관 질환 사망자에 유의한 영향을 미치는 것으로 나타났다.

$$\underset{}{Y_t} = \underset{(46.4817)}{100.9264} + \underset{(0.0773)}{0.2239}\,Y_{t-1} + \underset{(0.0791)}{0.1777}\,Y_{t-2} - \underset{(0.0765)}{0.1198}\,Y_{t-3} + \underset{(0.0717)}{0.5159}\,Y_{t-12} \qquad (2.4)$$

$$+ \underset{(2437)}{5481}X1_{t-1} - \underset{(2287)}{6614}X1_{t-2}.$$

표 2.13 서울시 월별 뇌혈관 질환 사망자에 대한 ADL 최종모형

Ordinary Least Squares Estimates

SSE	114467.168	DFE	99
MSE	1156	Root MSE	34.00344
SBC	1073.82709	AIC	1055.18302
MAE	26.4111186	AICC	1056.32587
MAPE	5.76192465	Regress R-Square	0.5928
		Total R-Square	0.5928

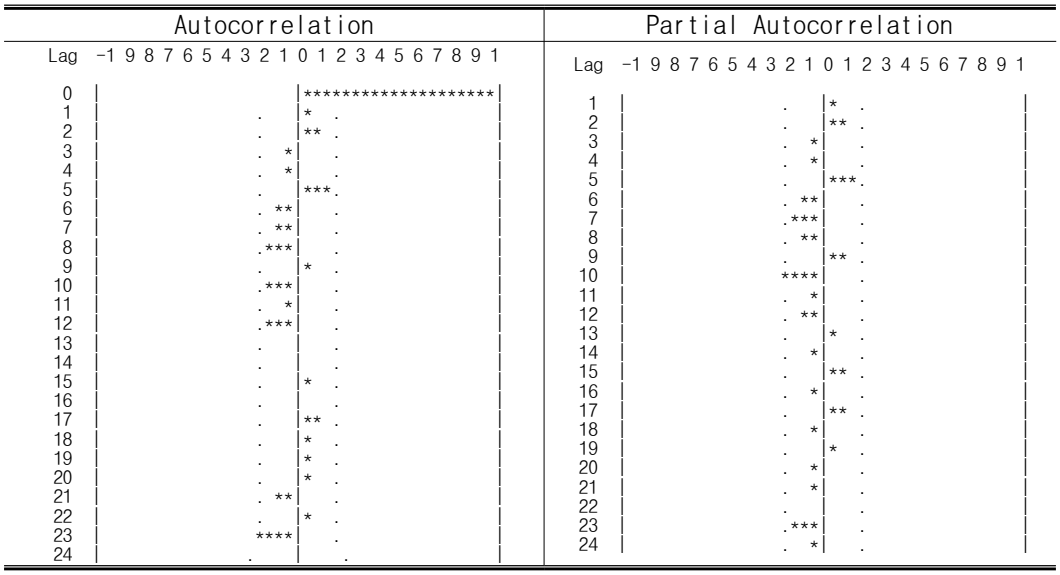
Miscellaneous Statistics

Statistic	Value	Prob	Label
Durbin's t	0.9903	0.1623	Pr > t

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	100.9264	46.4817	2.17	0.0323
y_1	1	0.2239	0.0773	2.90	0.0047
y_2	1	0.1777	0.0791	2.25	0.0269
y_3	1	-0.1198	0.0765	-1.57	0.1206
y_12	1	0.5159	0.0717	7.20	<.0001
x1_1	1	5481	2437	2.25	0.0267
x1_2	1	-6614	2287	-2.89	0.0047

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.41	6	0.6217	0.058	0.104	-0.057	-0.029	0.138	-0.082
12	12.04	12	0.4428	-0.094	-0.141	0.058	-0.152	-0.058	-0.150
18	13.36	18	0.7698	0.022	0.010	0.064	-0.016	0.076	0.066
24	17.70	24	0.8174	0.058	0.027	-0.108	0.059	-0.184	-0.009



2.3 65세 이상 서울시 월별 뇌혈관 질환 사망자 모형

65세 이상 연령층의 경우 작은 환경의 변화에도 민감하게 반응하는 연령층으로, 65세 이상 고령인구는 계속적으로 증가하는 추세에 있어 최근 고령인구에 대한 관심이 높아지고 있다. 따라서 이 연령층만을 대상으로 분석을 실시하였다.

2.3.1 65세 이상 뇌혈관 질환 사망자 분해모형

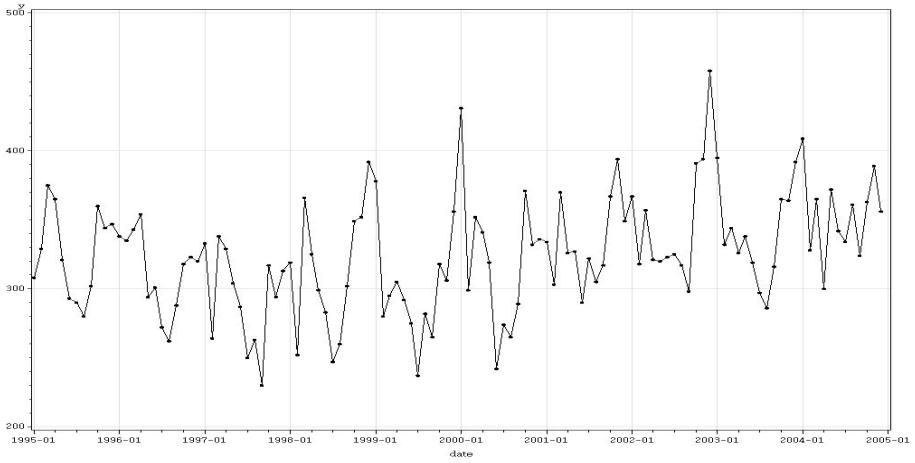


그림 2.4 65세 이상 뇌혈관 질환 사망자

그림 2.4의 65세 이상 서울시 월별 뇌혈관 질환 사망자는 추세가 일정하지 않고 선형적으로 증가하는 형태를 보이고 있다. 이는 한국사회가 '고령화 사회'로 진입하여 '고령사회'로 급속히 진입하고 있음을 보여주는 것이다. '고령화 사회'란 총 인구 중 65세 이상 인구가 총인구를 차지하는 비율이 7% 이상인 사회를 일컫는 것이고, '고령사회'는 총 인구 중 65세 이상 인구가 총 인구를 차지하는 비율이 14%이상인 사회를 일컫는 것으로 한국은 2000년에 65세 이상 노인인구가 약 7.2%로 이미 '고령화 사회'로 진입하였다. 아래 표 2.14는 2000년부터 2004년까지 통계청의 주민등록 인구통계의 65세 이상 노인인구로 계속 증가하고 있는 추세를 나타내고 있다.

표 2.14 주민등록 인구통계(65세 이상)

		2000년	2001년	2002년	2003년	2004년
전국	남	1,257,893	1,337,711	1,419,557	1,507,678	1,608,017
	녀	2,097,721	2,195,759	2,293,069	2,397,306	2,516,929
	전체	3,355,614	3,533,470	3,712,626	3,904,984	4,124,946
서울	남	212,862	227,039	241,134	258,813	280,994
	녀	345,704	358,858	371,649	388,096	408,992
	전체	558,566	585,897	612,783	646,909	689,986

증가 추세에 있는 65세 이상의 사망자 중 서울시 월별 뇌혈관 질환 사망자를 2.2.1절과 같은 방법으로 분해모형에 적합하였다. 분해모형 추정결과 오차항들의 자기상관을 확인 할 수 있는 DW통계량이 1.1786이고 p<0.0001로 자기상관이 존재하는 것으로 나타나 자기회귀오차를 가지는 회귀모형을 고려하였다. 또한 65세 이상 사망자의 경우 추세성분 t가 유의하게 나타나 이를 포함한 자기회귀오차를 가지는 회귀모형에 적합 시켜 추정하였다. 추정결과가 표 2.15에 주어져 있다. 자기회귀오차를 가지는 회귀모형의 추정결과 DW통계량이 2.0815이고 p-값이 0.6613로 더 이상 자기상관이 존재 하지 않음을 확인할 수 있었으며 추세성분 t가 유의한 것으로 나타났다. 포토맨토우 검정과 잔차의 SACF와 SPACF 역시 모형이 적합함을 보이고 있으므로 식(2.5)를 65세 이상 서울시 월별 뇌혈관 질환 사망자의 최종 자기회귀 오차를 가지는 회귀모형으로 선택하였다.

$$\begin{aligned}\hat{Y}_t &= 333.8935 + 0.4058t + 3.9104ID_{t,jan} - 53.0544ID_{t,feb} - 6.7004ID_{t,mar} \\ &\quad - 28.3013ID_{t,apr} - 39.2656ID_{t,may} - 62.7572ID_{t,jun} - 73.8640ID_{t,jul} \\ &\quad - 70.9866ID_{t,aug} - 66.4398ID_{t,sep} - 8.1650ID_{t,oct} - 11.5660ID_{t,nov} + \hat{e}_t \\ \hat{e}_t &= 0.4056\hat{e}_{t-1} \ . \end{aligned} \tag{2.5}$$

표 2.15 65세 이상 뇌혈관 질환 사망자의 자기회귀오차를 가지는 회귀모형

Maximum Likelihood Estimates

SSE	68380.8857	DFE	106
MSE	645.10270	Root MSE	25.39887
SBC	1169.19272	AIC	1130.16783
MAE	18.6017505	AICC	1134.16783
MAPE	5.80048488	Regress R-Square	0.5551
Durbin-Watson	2.0815	Total R-Square	0.6664

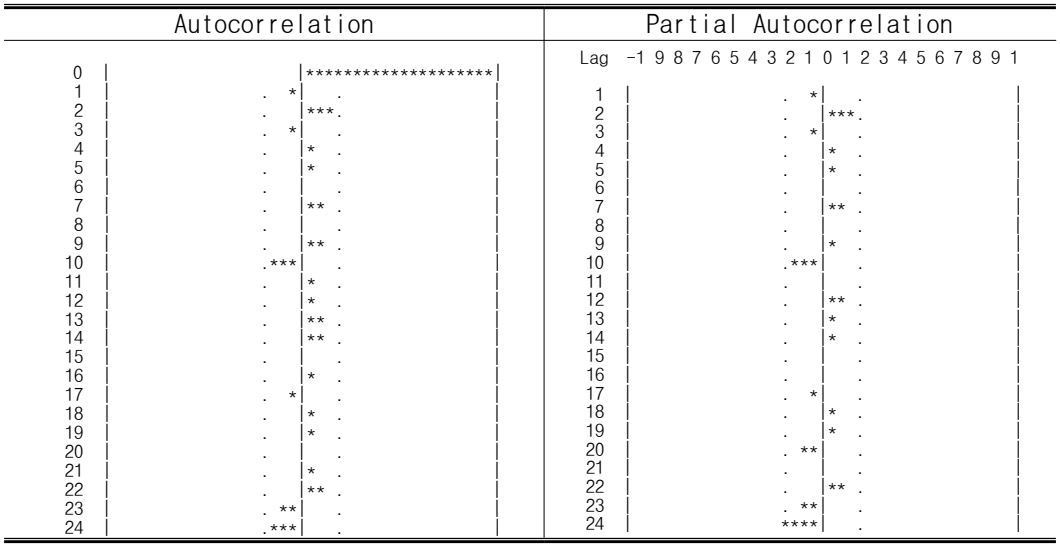
Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	2.0815	0.6613	0.3387

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	333.8935	11.2396	29.71	<.0001
t	1	0.4058	0.1113	3.65	0.0004
jan	1	3.9104	9.8107	0.40	0.6910
feb	1	-53.0544	11.4366	-4.64	<.0001
mar	1	-6.7004	12.0183	-0.56	0.5784
apr	1	-28.3013	12.2391	-2.31	0.0227
may	1	-39.2656	12.3189	-3.19	0.0019
jun	1	-62.7572	12.3354	-5.09	<.0001
jul	1	-73.8640	12.3069	-6.00	<.0001
aug	1	-70.9866	12.2109	-5.81	<.0001
sep	1	-66.4398	11.9624	-5.55	<.0001
oct	1	-8.1650	11.3230	-0.72	0.4724
nov	1	-11.5660	9.5595	-1.21	0.2290
AR1	1	-0.4056	0.0892	-4.55	<.0001

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.65	6	0.7235	-0.054	0.139	-0.039	0.055	0.051	0.003
12	8.68	12	0.7296	0.087	0.010	0.087	-0.131	0.059	0.047
18	12.03	18	0.8455	0.086	0.078	-0.005	0.056	-0.049	0.070
24	17.56	24	0.8236	0.039	-0.022	0.033	0.078	-0.082	-0.144



2.3.2 65세 이상 뇌혈관 질환 사망자 ADL모형

2.3.2.1 65세 이상 뇌혈관 질환 사망자 자기회귀모형

표 2.16에 나타난 바와 같이 65세 이상 서울시 월별 뇌혈관 질환 사망자의 SPACF가 시차 1과 12에서 유의하며 SACF는 12의 배수에 해당하는 시차를 따라 감소하고 시차 1, 2, 3,…에 대해서도 감소하는 패턴을 보이므로 $(1-\phi_1 B)(1-\Phi_1 B^{12})Y_t = \theta_0 + a_t$모형이 가능한 반면, SACF는 감소하는 패턴이고 SPACF는 시차 1, 12에서 유의한 것으로 간주하여 $(1-\phi_1 B-\phi_{12} B^{12})Y_t = \theta_0 + a_t$모형도 적합할 수도 있다고 판단되어 두 모형을 적합 시켰다.

표2.16 65세 이상 뇌혈관 질환 사망자 SACF와 SPACF

Name of Variable = y

Mean of Working Series 324
Standard Deviation 41.32715
Number of Observations 120

Autocorrelation

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Erro
0	1707.933	1.00000	|********************|	0
1	870.708	0.50980	|**********	0.091287
2	576.208	0.33737	|*******	0.112539
3	160.467	0.09395	|** .	0.120673
4	-152.650	-.08938	. **| .	0.121281
5	-92.100000	-.05392	. *| .	0.121829
6	-193.792	-.11347	. **| .	0.122027
7	-25.833333	-.01513	. | .	0.122903
8	-114.192	-.06686	. *| .	0.122919
9	202.033	0.11829	. |** .	0.123222
10	384.400	0.22507	. |*****	0.124164
11	627.775	0.36756	. |*******	0.127519
12	1057.517	0.61918	. |************	0.136062
13	671.817	0.39335	. |********	0.157805
14	441.200	0.25832	. |***** .	0.165775
15	89.783333	0.05257	. |* .	0.169096
16	-223.767	-.13102	. ***| .	0.169232
17	-155.225	-.09088	. **| .	0.170075
18	-165.475	-.09689	. **| .	0.170479
19	-4.416667	-.00259	. | .	0.170938
20	-76.983333	-.04507	. *| .	0.170938
21	195.683	0.11457	. |** .	0.171037
22	405.492	0.23742	. |***** .	0.171675
23	427.008	0.25001	. |***** .	0.174390
24	780.783	0.45715	. |*********	0.177352

Partial Autocorrelation

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.50980	. |**********
2	0.10468	. |** .
3	-0.15492	.***| .
4	-0.15954	.***| .
5	0.10795	. |** .
6	-0.06511	. *| .
7	0.06463	. |* .
8	-0.10625	. **| .
9	0.24532	. |*****
10	0.15039	. |***.
11	0.23040	. |*****
12	0.42699	. |*********
13	-0.08794	. **| .
14	-0.10052	. **| .
15	0.01664	. | .
16	-0.10304	. **| .
17	0.07257	. |* .
18	0.02348	. | .
19	0.00364	. | .
20	-0.07560	. **| .
21	0.07006	. |* .
22	0.14149	. |***.
23	-0.13689	.***| .
24	0.09275	. |** .

두 모형의 적합 결과 AIC와 SBC통계량을 기준으로 비교하였을 때 각각 두 모형의 AIC통계량이 1156.735, 1155.619이고, SBC통계량이 1165.098, 1163.981로 큰 차이는 없지만 해석의 용이성을 위해서 모형 $(1-\phi_1 B-\phi_{12}B^{12})Y_t=\theta_0+a_t$를 고려한다. 모형 $(1-\phi_1 B-\phi_{12}B^{12})Y_t=\theta_0+a_t$을 최우추정법으로 추정한 결과가 표 2.17에 주어져 있다. 모수추정치는 모두 유의하게 나타났으며 2.3.1절의 자기회귀오차를 가지는 회귀모형과 비교 시 AIC통계량이 1130.198에서 1155.619로 증가한 반면 SBC통계량이 1169.193에서 1163.981로 감소하였다. 그러나 포토맨토우 검정과 잔차의 SACF와 SPACF가 모형이 적합함을 보이고 있으므로 식 (2.6)을 65세 이상 서울시 뇌혈관 질환 사망자수의 월별 사망

자의 최종 자기회귀모형으로 선택하였다.

$$Y_t = \underset{(13.01388)}{330.60146} + \underset{(0.06690)}{0.29886}\,Y_{t-1} + \underset{(0.07011)}{0.54647}\,Y_{t-12}. \qquad (2.6)$$

표 2.17 65세 이상 뇌혈관 질환 사망자에 대한 자기회귀모형

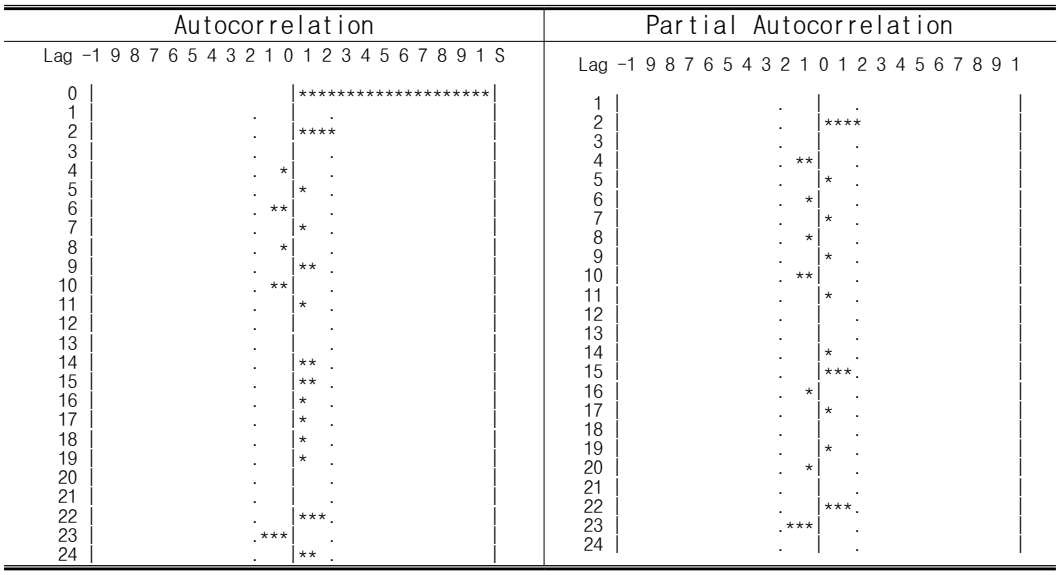
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	330.60146	13.01388	25.40	<.0001	0
AR1,1	0.29886	0.06690	4.47	<.0001	1
AR1,2	0.54647	0.07011	7.79	<.0001	12

Constant Estimate	51.13405
Variance Estimate	832.0829
Std Error Estimate	28.84585
AIC	1155.619
SBC	1163.981
Number of Residuals	120

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	7.14	4	0.1286	0.013	0.214	-0.002	-0.041	0.026	-0.093
12	10.73	10	0.3793	0.068	-0.071	0.082	-0.086	0.054	-0.023
18	13.79	16	0.6144	0.007	0.076	0.109	0.027	0.054	0.027
24	21.93	22	0.4638	0.045	-0.011	0.001	0.157	-0.146	0.078



2.3.2.2 65세 이상 뇌혈관 질환 사망자 ADL모형

모형의 설명력 및 예측력을 높이기 위해 자기회귀모형 (2.6)에 추가적인 설명변수를 고려한 ADL모형을 살펴보겠다.

65세 이상 서울시 월별 뇌혈관 질환 사망자수의 경우 2.2.2절과 동일한 방법으로 그랜져 인과성검정을 실시한 결과 $X1(SO_2)$만이 유의한 것으로 나타났다. 즉 $X1$만이 사망자의 과거 값에 포착되지 않은 부분이외에 65세 이상 서울시 월별 뇌혈관 질환 사망자를 예측하는데 유용한 정보를 포함하고 있는 것으로 보인다. 따라서 2.3.2.1절의 자기회귀모형에 $X1$의 1-5차까지의 시차변수를 포함한 ADL모형을 적합 후 유의하지 않은 변수를 제거하여 최종모형을 선택하였다. 65세 이상 서울시 월별 뇌혈관 질환 사망자의 ADL최종모형 추정결과가 표 2.18에 주어져 있다. 추정결과 한달 전과 두달 전의 아황산가스($X1$)농도가 65세 이상 서울시 월별 뇌혈관 질환 사망자에 유의한

영향을 미치는 것으로 나타나 전체 사망자수와 유사한 결과임을 확인 할 수 있다. 최종으로 추정된 65세 이상 서울시 월별 뇌혈관 질환 사망자의 ADL 모형은 식 (2.7)과 같다.

$$Y_t = \underset{(26.3419)}{75.7119} + \underset{(0.0697)}{0.2936}\,Y_{t-1} + \underset{(0.0746)}{0.5111}\,Y_{t-12} + \underset{(1899)}{5204}X1_{t-1} - \underset{(1799)}{6531}X1_{t-2}. \qquad (2.7)$$

표 2.18 65세 이상 뇌혈관 질환 사망자에 대한 ADL 최종모형

Ordinary Least Squares Estimates

SSE	79087.7819	DFE	101
MSE	783.04735	Root MSE	27.98298
SBC	1025.30887	AIC	1011.99167
MAE	21.9646511	AICC	1012.59167
MAPE	6.88090393	Regress R-Square	0.5891
		Total R-Square	0.5891

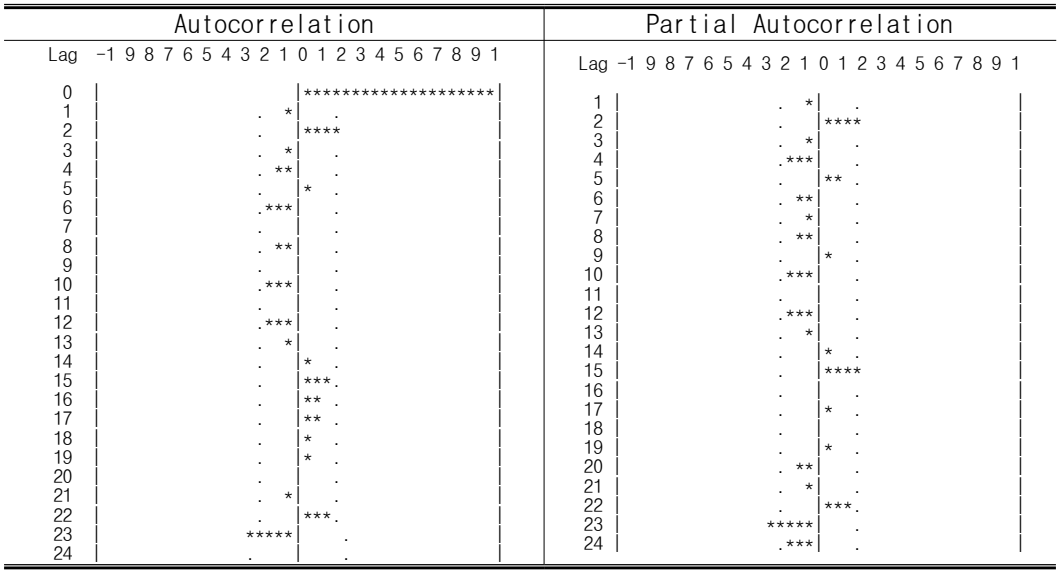
Miscellaneous Statistics

Statistic	Value	Prob	Label
Durbin's t	-0.3949	0.3469	Pr > t

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	75.7119	26.3419	2.87	0.0049
y_1	1	0.2936	0.0697	4.21	<.0001
y_12	1	0.5111	0.0746	6.85	<.0001
x1_1	1	5204	1899	2.74	0.0072
x1_2	1	-6531	1799	-3.63	0.0004

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	6.92	6	0.3281	-0.027	0.190	-0.044	-0.089	0.071	-0.131
12	13.16	12	0.3573	0.015	-0.119	0.022	-0.152	0.002	-0.170
18	16.62	18	0.5495	-0.063	0.029	0.133	0.095	0.078	0.036
24	24.35	24	0.4419	0.051	-0.015	-0.055	0.131	-0.267	-0.025


Autocorrelation
Lag -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0 |********************|
1 | . *| . |
2 | . |**** |
3 | . *| . |
4 | . **| . |
5 | . |* . |
6 | .***| . |
7 | . | . |
8 | . **| . |
9 | . | . |
10 | .***| . |
11 | . | . |
12 | .***| . |
13 | . *| . |
14 | . |* . |
15 | . |***. |
16 | . |** . |
17 | . |** . |
18 | . |* . |
19 | . |* . |
20 | . | . |
21 | . *| . |
22 | . |***. |
23 | *****| . |
24 | . | . |
Partial Autocorrelation
Lag -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1 | . *| . |
2 | . |**** |
3 | . *| . |
4 | .***| . |
5 | . |** . |
6 | . **| . |
7 | . *| . |
8 | . **| . |
9 | . |* . |
10 | .***| . |
11 | . | . |
12 | .***| . |
13 | . *| . |
14 | . |* . |
15 | . |**** |
16 | . | . |
17 | . |* . |
18 | . | . |
19 | . |* . |
20 | . **| . |
21 | . *| . |
22 | . |***. |
23 | *****| . |
24 | .***| . |

제3장 음이항 회귀모형

1.1절에서 언급했던 바와 같이 기존 연구에서 '일반적으로 일별 사망자수는 전체 인구집단으로 볼 때 매우 드물게 발생하는 사건이고 단위가 발생 숫자이기 때문에 항상 양의 정수가 된다.'하여 포아송 회귀모형에 적합하여 분석을 실시하였다. 그러나 앞에서 서울시 일별 호흡기 질환 사망자의 경우 포아송 분포를 따르지 않음을 확인했으며, 본 논문에서는 일별 사망자가 아닌 월별 사망자를 대상으로 분석을 실시하였으므로 종속변수인 월별 사망자가 포아송분포가 따른다는 가정은 적합하지 않은 것으로 보인다. 따라서 서울시 월별 뇌혈관 질환 사망수자가 포아송 법칙에 좀 더 유연한 음이항 분포를 따른다고 가정하고 음이항 회귀모형에 적합하여 분석을 실시 하고자 한다.

분석에 앞서 콜모고로프-스미르노프 검정(Kolmogorov-Smirnov Test)을 통하여 종속변수인 서울시 월별 뇌혈관 질환 사망자가 음이항 분포를 따르는 지를 검토한 후 음이항 회귀모형의 설명을 돕기 위해 포아송 회귀모형을 먼저 살펴보겠다.

3.1 콜모고로프-스미르노프 검정

콜모고로프-스미르노프는 적합도 검정을 위한 방법으로 함수들 사이의 최대 수직거리를 이용하여 함수들이 얼마나 유사한지를 측정하는 통계적 방법을 개발하였다. 콜모고로프-스미르노프 통계량(Kolmogorov-Smirnov

Statistics)은 경험적 분포와 귀무가설에서 규정한 분포사이의 최대수직거리의 함수를 고려하는 콜모고로프 타입 통계량과, 두 경험적 분포함수사이의 최대수직거리의 함수를 고려하는 스미르노프 타입 통계량이 있다(Conover, 1980).

적합도 검정으로 많이 사용되는 카이제곱검정이 표본의 크기가 작은 경우 정확한 검정을 수행하지 못하는 것과 달리 콜모고로프-스미르노프 검정(Kolmogorov-Smirnov Test)은 소표본인 경우에도 정확한 검정결과를 제공한다. 그러나 이 검정방법은 귀무가설에서 규정한 분포가 연속형인 경우 사용하는 검정 법으로 이산형인 경우에도 이 검정방법이 적용 가능하지만 이 경우에는 진실 된 유의수준이 α이하가 되어 검정결과는 좀 더 보수적이게 된다(심정욱 등, 2003).

콜모고로프-스미르노프 검정 방법은 다음과 같다.

$$H_0 : F(x) = F^*(x), \qquad -\infty \leq x \leq \infty,$$

$$H_1 : F(x) \neq F^*(x), \qquad at\ laest\ one\ value\ of\ x.$$

위 가설 검정을 위한 검정 통계량은 T로 표기하고 $F^*(x)$와 $S(x)$사이의 최대수직 거리로 다음과 같이 정의한다.

$$T = \sup|F^*(x) - S(x)| \ ,$$

여기서

$F^*(x)$: 귀무가설에서 규정한 분포,

$S(x) = \frac{[x\text{보다 작거나 같은 관측값 } X_i\text{의 개수}]}{n}$ (전명식,2007).

위 검정통계량 T와 α하에서의 $1-\alpha$분위수 $w_{1-\alpha}$(quantiles of the Kolmogorov Test Statistic)값을 비교하여 검정통계량 T값이 $w_{1-\alpha}$값을 초과하면 귀무가설을 기각한다. $w_{1-\alpha}$값은 양측검정인 경우 $n \leq 20$일 때 정확한 검정결과를 제공한다. 직관적으로 볼 때 경험적 분포 $S(x)$가 귀무가설에서 규정한 분포 $F^*(x)$에 충분히 가까우면, 즉 검정통계량 값이 작을 경우 귀무가설 H_0를 채택하게 된다.

본 논문의 경우 앞에서 언급했던 바와 같이 서울시 월별 뇌혈관 질환 사망자(Y_t)가 음이항 분포를 따른다는 가설을 검정하고자 한다. 앞에서 양측검정의 경우 $n \leq 20$일 때 정확한 검정결과를 제공한다 하였으므로 20개의 표본을 임의 추출하여 검정을 실시하였다. 유의수준 α=0.05하에서 $n=20$인 경우 $w_{1-\alpha}$값은 0.294로 기각역은 $\{T: T \geq 0.294\}$이다. 통계 소프트웨어에서 제공되는 규정된 분포의 CDF값과 경험적 분포의 누적 백분율 값(cumulative percentile)을 비교한다.

$$\begin{aligned} T &= \sup|F^*(x) - S(x)| \\ &= |F^*(479) - S(479)| \\ &= 0.1887 \end{aligned}$$

검정 통계량 T=0.1887은 α=0.05하에서 임계치 0.294보다 작으므로 귀무가설을 기각 할 수 없다. 즉, 서울시 월별 뇌혈관 질환 사망자가 음이항 분포를 따름을 확인할 수 있다. 표 3.1은 서울시 월별 뇌혈관 질환 사망자의 콜모고로프-

스미르노프 검정 결과 이다. 또한 그림 3.1의 규정된 분포와 경험적 분포의 CDF 플롯(plot) 결과 역시 검정결과를 뒷 받침해주고 있다.

표3.1 콜모고로프 검정

Y_t	규정된 분포$[F^*(x)]$	경험적 분포$[S(x)]$	$T=\|F(x)^* - S(x)\|$
370	0.0300	0.0000	0.0300
386	0.0630	0.0500	0.0130
395	0.0907	0.1000	0.0093
404	0.1261	0.1500	0.0239
411	0.1591	0.2000	0.0409
420	0.2085	0.2500	0.0415
438	0.3279	0.3000	0.0279
439	0.3351	0.3500	0.0149
452	0.4335	0.4000	0.0335
479	**0.6387**	**0.4500**	**0.1887**
484	0.6735	0.5000	0.1735
485	0.6803	0.5500	0.1303
491	0.7195	0.6000	0.1195
503	0.7894	0.6500	0.1394
505	0.7999	0.7000	0.0999
508	0.8150	0.7500	0.0650
517	0.8556	0.8000	0.0556
529	0.8992	0.8500	0.0492
542	0.9342	0.9000	0.0342
621	0.9978	0.9500	0.0478

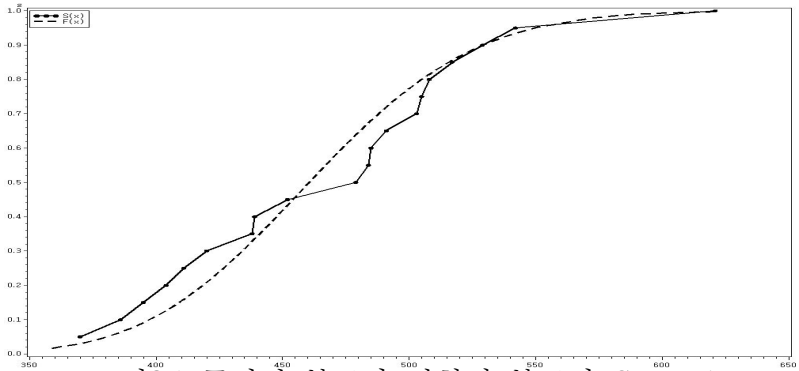


그림3.1 규정된 분포와 경험적 분포의 CDF Plot

본 논문의 경우 귀무가설에서 규정한 분포는 음이항 분포로 이산형 분포이므로 언급했던 바와 같이 이 경우 참(true) 유의수준이 α보다 작다. 그러나 유의수준이 작아질수록 임계치 $w_{1-\alpha}$가 커져 앞의 검정결과는 변함이 없다.

3.2 포아송 회귀모형

포아송 회귀모형은 도수자료 분석을 위한 모형에 널리 사용된다. 이 모형은 공변량에 의해서 결정되는 모수 μ를 가지는 포아송 분포로부터 유도된다. 모형에서 종속변수 Y_i는 관심 있는 사건의 수이고, X_i는 Y_i를 결정하는 선형독립 공변량 벡터이다. 회귀모형은 $E(Y_i|X_i)=\mu(X_i,\beta)$와 같이 모수 β와 함수 $\mu(X_i,\beta)$ 그리고 공변량 $X_i^{'}=[X_{1,i},X_{2,i},\cdots X_{k,i}]$의 k차원 벡터가 주어 졌을 때 Y_i의 조건부분포에 의해 결정된다(Cameron, 1998).

즉 X_i가 주어졌을 때 Y_i는 다음과 같은 확률밀도함수를 가지는 포아송 분포를 따른다.

$$f(Y_i|X_i)=\frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!}\ , \quad Y_i=0,1,2\ldots. \tag{3.1}$$

위 식 (3.1)의 모수 μ_i는 i번째 공변량들의 결합으로 구성되는데 μ_i는 0보다 작을 수 없으므로 포아송 회귀모형은 공변량의 로그선형함수로 식 (3.2)와 같이 표현된다(Paul, 1999).

$$\log[E(Y_i|X_i)]=X_i'\beta=X_{1i}\beta_1+X_{2i}\beta_2+\cdots+X_{ki}\beta_k. \tag{3.2}$$

위와 같은 로그선형모형의 표현에서 평균모수는 다음과 같이 표현된다.

$$\mu_i = \exp(X_i^{'}\beta)\,, \qquad \mu_i > 0. \tag{3.3}$$

식 (3.1)과 (3.3)에 의해 포아송 회귀모형을 정의한다. 식 (3.2)에 나타난 것처럼 로그선형모형의 평균모수는 다음과 같이 승법 형태로 표현된다.

$$\begin{aligned} E[Y_i|X_i] &= \exp(X_i^{'}\beta) \\ &= \exp(X_{i1}\beta_1)\cdot\exp(X_{2i}\beta_2)\cdots\exp(X_{ki}\beta_k). \end{aligned}$$

만약 가법형태$\{E[Y_i|X_i] = X_i^{'}\beta = \sum_{j=1}^{k} X_{ji}\beta_i\}$로 표현된다면 μ_i가 음수가 아니라는 가정에 위배될 수 있으므로 가법형태의 표현은 포아송 회귀모형에 적절하지 못하다. 앞에서 언급한 바와 같이 이러한 이유로 승법형태의 표현을 위해 연결함수로 로그함수를 사용한다.

식 (3.3)의 포아송 회귀모형의 조건부 평균함수에 관측되지 않은 오차항 τ_i가 존재하여 조건부 평균함수가 $E[Y_i|X_i,\tau_i]$로 표현된다면 Y_i의 주변분포(marginal distribution)는 τ_i분포의 적률(moment)을 포함하게 되고 이를 이용해 혼합 포아송분포(mixed Poisson distribution)를 유도한다. 다음 절에서 혼합 포아송분포에 대해 살펴보겠다.

3.3 이질성과 과대산포

3.1절에서 언급했던 바와 같이 포아송 회귀모형은 '평균과 분산이 같다'는 포아송 법칙을 가정하지만 실제자료의 경우 과대산포 문제가 많이 발생하여

포아송 회귀모형은 실제자료에 적합 되는 경우는 적다(권민경, 2003). 음이항 회귀모형은 이러한 과대산포를 해결하기 위한 방법 중 하나로써 모형에 과대산포 모수를 포함하는 모형이다. 따라서 음이항 회귀모형에 대한 설명에 앞서 과대산포를 유도하는 이질성(heterogeneity)항과 과대산포에 대하여 살펴보고 3.3.1절에서 이질성항을 포함함으로서 과대산포를 모형에 반영한 음이항 회귀모형에 대해 알아보겠다.

이질성항이 없는 포아송 회귀모형에서 $(Y_i|X_i)$는 관측 가능한 X_i가 주어진 상황에서의 분포를 의미한다. 즉 조건부 평균함수가 X_i의 비확률(nonstochastic)함수로 정의됨을 의미한다. 혼합모형(mixture model)의 경우 $(Y_i|X_i,\tau_i)$의 분포를 지정하는데 여기서 τ_i는 i번째 관측치의 관측되지 않은 이질성항이다. 즉, 개체들이 관측된 공변량에 의해 충분히 설명되지 못해 각각의 개체들이 확률변수 τ_i에 의해 확률적으로 다르다고 가정한다.

확률 항(stochastic term) τ_i를 고려해 준 모형에서는 Y_i와 (X_i,τ_i)를 연결해주는 정확한 함수형태를 지정해 주어야 하는데 일반적으로 다음과 같은 함수형태를 가진다.

$$E[Y_i|X_i,\tau_i] = \exp(X_i'\beta)\cdot\tau_i,$$

여기서 확률 항 τ_i는 공변량과 독립이다. 위와 같은 승법형태의 이질성항의 가정은 수리적으로 계산이 편리하고 앞에서 언급했던 바와 같이 Y_i가 음수가 아니라는 제약을 위반할 수 있는 가법형태 보다 유용하다.

이질성항이 승법형태로 모형안에 포함된 혼합모형의 조건부 분산은 고유의

포아송모형의 조건부 분산보다 크다. 이는 포아송 모형이 모형에서 관측되지 않은 이질성을 무시한 결과로써 과대산포의 일반적인 해석에 기초가 된다. 혼합모형의 조건부 평균은 $\mu_i = \exp(X_i'\beta)$ 대신 다음과 같은 표현으로 대체한다.

$$\mu_i^* = E[Y_i|\mu_i, \tau_i] = \mu_i \tau_i, \tag{3.4}$$

여기서 $\tau_i = \exp(\epsilon_i)$는 관측되지 않고 제거된 오차를 반영할 수 있으며 알려진 모수를 갖는 iid분포를 따르고 X_i와 독립임을 가정한다. 편의를 위해 τ_i는 $E[\tau_i] = 1$와 $Var[\tau_i] = \sigma_\tau^2$를 갖는 iid분포를 따른다고 가정한다. 만약 $E[y_i|X_i, \tau_i] = Var[y_i|X_i, \tau_i] = \mu_i$라 가정한다면, 이는 포아송 분포임을 의미하는 것이다.

Y_i의 적률(moment)은 다음과 같이 유도된다.

$$E[Y_i|X_i] = \mu_i, \tag{3.5}$$

$$Var[Y_i|X_i] = \mu_i[1 + \sigma_\tau^2 \mu_i] > E[Y_i|X_i]. \tag{3.6}$$

$f(Y_i|X_i, \tau_i)$를 식(3.4)과 같이 μ_i가 μ_i^*로 대체된 식에 의해 얻어진 확률함수라 정의하고, $g(\tau_i)$는 τ_i의 확률밀도 함수라 정의하면 $(Y|X)$의 혼합 주변밀도함수(mixed marginal density)는 아래와 같이 $f(Y_i|X_i, \tau_i)$를 τ_i로 적분함으로써 얻어진다.

$$f(Y|X) = \int f(Y|X, \tau) \cdot g(\tau) d\tau. \tag{3.7}$$

비록 이 혼합 포아송분포의 정확한 형태는 $g(\tau_i)$의 선택에 의존하지만 과대산포의 일반적인 특성은 $g(\tau_i)$에 의존하지 않는다. 위 (3.7)식을 이용하여 3.3.1절에서 음이항 분포를 유도하겠다.

3.3.1 음이항 회귀모형

앞에서 설명했던 혼합 포아송모형 중 포아송-감마 혼합모형으로부터 음이항모형이 유도되고 해석되어 진다. 확률도수 Y_i가 조건부 포아송 분포를 따르고 다음과 같은 확률밀도 함수를 가진다고 가정한다.

$$f(Y_i|\mu_i^*) = \frac{\exp(-\mu_i^*) \cdot \mu_i^{*Y_i}}{Y_i!} \qquad Y_i = 0,1,\ldots, \tag{3.8}$$

여기서 모수 μ_i^*는 랜덤 항을 포함하며, 오차항인 이 랜덤 항은 승법형태로 조건부 평균에 곱해져 모형 안에 들어온다. 즉, 다음과 같이 표현된다.

$$\begin{aligned} \mu_i^* &= \exp(\beta_0 + X_i'\beta_1 + \epsilon_i) \\ &= e^{(\beta_0 + X_i'\beta_1)} e^{\epsilon_i} \\ &= \mu_i \tau_i, \end{aligned} \tag{3.9}$$

여기서 $\exp(\beta_0 + \epsilon_i)$는 확률 절편 항으로 해석되고, $\mu_i = e^{(\beta_0 + X_i'\beta_1)}$ 이고 $\tau_i = e^{\epsilon_i}$이다.

따라서 식 (3.8)을 다시 표현하면 다음과 같다.

$$f(Y_i|X_i,\tau_i)=\frac{\exp(-\mu_i\tau_i)(\mu_i\tau_i)^{Y_i}}{Y_i!}.$$

Y의 조건부 주변분포는 위 식을 τ_i로 적분함으로써 구할 수 있다.

$$\begin{aligned} f(Y_i|X_i) &= \int f(Y_i|X_i,\tau_i)g(\tau_i)d\tau_i \\ &\equiv E_\tau[f(Y_i|X_i,\tau_i)], \end{aligned} \tag{3.10}$$

여기서 $f(\bullet)$와 $g(\bullet)$를 각각 포아송 분포와 감마분포를 따른다고 가정하면 적분의 해는 명확해진다. τ_i가 감마(gamma)분포 $g(\tau,\theta,1/\theta)$를 따른다고 가정한다.

$$g(\tau_i,\theta,1/\theta)=\frac{\theta^\theta}{\Gamma(\theta)}\tau_i^{\theta-1}e^{-(\tau_i\theta)}, \qquad \theta>0, \tag{3.11}$$

여기서 $E[\tau_i]=\theta\cdot\frac{1}{\theta}=1$이고, $Var[\tau_i]=\theta\cdot(\frac{1}{\theta})^2=\frac{1}{\theta}$이다.

감마함수에 대한 정의를 이용하여 식 (3.10)을 유도하면 다음과 같다.

$$f(Y_i|X_i) = \int_0^{\infty} f(Y_i|X_i,\tau_i)g(\tau_i)d\tau_i \tag{3.12}$$

$$= \frac{\theta^{\theta}\mu_i^{y_i}}{Y_i!\Gamma(\theta)} \int_0^{\infty} e^{-(\mu_i+\theta)\tau_i} \tau_i^{\theta+Y_i-1} d\tau_i$$

$$= \frac{\theta^{\theta}\mu_i^{Y_i}\Gamma(Y_i+\theta)}{Y_i!\Gamma(\theta)(\theta+\mu_i)^{\theta+Y_i}}$$

$$= \frac{\Gamma(Y_i+\theta)}{Y_i!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} \left(\frac{\mu_i}{\theta+\mu_i}\right)^{Y_i},$$

여기서 $\gamma = \frac{1}{\theta}$라 놓으면 Y_i의 주변분포는 다음과 같은 음이항분포의 확률밀도 함수가 된다.

$$f(Y_i|X_i) = \frac{\Gamma(Y_i+\gamma^{-1})}{Y_i!\Gamma(\gamma^{-1})} \left(\frac{\gamma^{-1}}{\gamma^{-1}+\mu_i}\right)^{\gamma^{-1}} \left(\frac{\mu_i}{\gamma^{-1}+\mu_i}\right)^{Y_i}, \quad Y_i = 0,1,2,\ldots. \tag{3.13}$$

위와 같은 확률밀도 함수를 가지는 음이항 분포는 다음과 같은 조건부 평균과 조건부 분산을 가진다.

$$E[Y_i|X_i] = \mu_i,$$
$$Var[Y_i|X_i] = \mu_i(1+\gamma\mu_i) > \mu_i.$$

앞 포아송 회귀모형과 같이 $E[Y_i|X_i] = \mu_i$의 모수 μ_i는 i번째 공변량들의 결합

으로 구성되는데 μ_i는 0보다 작을 수 없으므로 음이항 회귀모형은 공변량의 로그선형함수로 식 (3.2)와 같이 표현된다.

$$\log[E(Y_i|X_i)] = X_i'\beta = X_{1i}\beta_1 + X_{2i}\beta_2 + \cdots + X_{ki}\beta_k. \tag{3.14}$$

이를 행렬로 표현하면 다음과 같이 표현된다.

$$\log[E(Y_i|X_i)] = \log[E(Y_1, Y_2, \cdots, Y_n|X_i)] = X_i'\beta = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

여기서 $Y_1, Y_2, \cdots, Y_n$은 서로 독립이고 음이항 분포를 따르는 확률변수로 $Y_i = [Y_1, Y_2, \cdots, Y_n]'$는 $n \times 1$벡터이다. $X_i = [X_{1i}, X_{2i}, \cdots, X_{ki}]$는 독립변수들로 이루어진 $n \times (k+1)$ 공변량 행렬이고, $\beta = [\beta_0, \beta_1, \cdots, \beta_k]'$는 $(k+1) \times 1$ 모수벡터이다.

음이항 회귀모형에서 모수 γ와 β는 최우추정법에 의해 추정될 수 있는데 분산이 평균의 2차 함수로 표현되는 음이항 회귀모형의 로그우도함수는 다음과 같이 표현된다.

$$L = \sum_{i=1}^{N} [\sum_{j=0}^{Y_i - 1} \ln(j + \gamma^{-1}) - \ln(Y_i!) - (Y_i + \gamma^{-1}) \ln(1 + \gamma \cdot \exp(X_i'\beta)) + Y_i \ln(\gamma) + Y_i X_i'\beta] \tag{3.15}$$

여기서 $\Gamma(Y+\gamma)/\Gamma(\gamma) = \prod_{j=0}^{Y-1}(j+\gamma)$이다.

γ와 β는 다음 식에 의해 추정된다.

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{N} \frac{Y_i - \mu_i}{1+\gamma\mu_i} \cdot X_i, \tag{3.16}$$

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{N} [-\gamma^{-2} \sum_{j=0}^{Y_i-1} \frac{1}{(j+\gamma^{-1})} + \gamma^{-2}\ln(1+\gamma\mu_i) + \frac{Y_i - \mu_i}{\gamma(1+\gamma\mu_i)}]. \tag{3.17}$$

최대우도 추정치 $(\hat{\beta},\hat{\gamma})$는 연속적인 반복에 의해 구할 수 있다. 먼저 γ의 초기값 $\gamma_{(0)}$를 사용하여 각각의 β에 대해 $L(\beta,\gamma)$를 최대화 시키는 $\beta_{(1)}$를 구한다. 그리고 β값을 $\beta_{(1)}$으로 고정시켜 놓고, 각각의 γ중 $L(\beta,\gamma)$를 최대화 하는 $\gamma_{(1)}$을 구한다. 이러한 'Newton-Raphson' 반복 방법에 의해 γ가 고정되어 있을 때와 β가 고정되어 있을 때의 과정을 순환, 반복하여 $L(\beta,\gamma)$를 최대로 하는 최대우도 추정치 $(\hat{\beta},\hat{\gamma})$를 구할 수 있다(Ismail and Jemain, 2007).

3.4 서울시 월별 뇌혈관 질환 사망자의 음이항 회귀모형

본 논문의 종속변수인 서울시 월별 뇌혈관 질환 사망자는 음이 아닌 정수로 3.1절에서 콜모고로프-스미르노프 검정을 통하여 음이항 분포를 따르는 것을 보였다. 따라서 음이항 회귀모형을 이용하여 대기오염물질 및 기상변수가 서울시 월별 뇌혈관 질환 사망자에 미치는 영향을 분석해보도록 하겠다. 음이항 회귀모형의 적합을 위해 모형에 가능한 변수를 모두 포함한 완전모형을 적합한 후 유의하지 않은 변수를 제거하여 최종모형을 선택하였다. 2.2.2절에서 모형에 포함하는 설명변수의 시차길이를 5로 결정하였으므로 음이항 회귀모형에

서도 X1부터 X6의 5차 시차까지 고려하였고, ADL모형과 달리 t시점의 설명변수도 모형에서 고려하였다. 음이항 회귀모형 적합에 일반적으로 사용되는 SAS의 'PROC GENMOD'를 이용하여 완전 모형을 적합 후 유의하지 않은 변수를 제거하여 최종모형을 선택하였다. 가능한 모든 변수를 포함시킨 서울시 월별 뇌혈관 질환 사망자의 음이항 회귀모형의 추정결과는 표 3.2와 같다.

표 3.2 서울시 월별 뇌혈관 질환 사망자에 대한 음이항 회귀모형

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	66	101.6215	1.5397
Scaled Deviance	66	101.6215	1.5397
Pearson Chi-Square	66	102.3495	1.5507
Scaled Pearson X2	66	102.3495	1.5507
Log Likelihood		243029.9467	
Full Log Likelihood		-488.3364	
AIC (smaller is better)		1052.6728	
AICC (smaller is better)		1098.9853	
BIC (smaller is better)		1152.7925	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	6.4547	0.3176	5.8322	7.0771	413.08	<.0001
x1	1	-12.5390	6.6714	-25.6147	0.5367	3.53	0.0602
x1_1	1	18.6269	8.9714	1.0432	36.2106	4.31	0.0379
x1_2	1	-19.3404	8.3321	-35.6709	-3.0099	5.39	0.0203
x1_3	1	17.3842	8.2809	1.1538	33.6145	4.41	0.0358
x1_4	1	-3.9738	7.3136	-18.3083	10.3607	0.30	0.5869
x1_5	1	-1.6344	4.8227	-11.0867	7.8179	0.11	0.7347
x2	1	-4.8197	3.8885	-12.4411	2.8017	1.54	0.2152
x2_1	1	-6.5652	3.8531	-14.1171	0.9867	2.90	0.0884
x2_2	1	-0.6947	3.7853	-8.1138	6.7244	0.03	0.8544
x2_3	1	6.4768	3.8370	-1.0435	13.9972	2.85	0.0914
x2_4	1	0.9734	3.8241	-6.5216	8.4685	0.06	0.7991
x2_5	1	-15.4177	3.7199	-22.7086	-8.1269	17.18	<.0001
x3	1	0.3120	3.1156	-5.7945	6.4186	0.01	0.9202
x3_1	1	3.5806	3.1941	-2.6798	9.8410	1.26	0.2623
x3_2	1	1.1756	3.2412	-5.1771	7.5283	0.13	0.7168
x3_3	1	-10.5887	2.9760	-16.4215	-4.7560	12.66	0.0004
x3_4	1	6.6086	2.9113	0.9024	12.3147	5.15	0.0232
x3_5	1	-2.8007	2.7495	-8.1896	2.5883	1.04	0.3084
x4	1	0.0006	0.0010	-0.0014	0.0026	0.30	0.5817
x4_1	1	-0.0008	0.0011	-0.0029	0.0013	0.54	0.4640
x4_2	1	-0.0001	0.0012	-0.0024	0.0022	0.01	0.9346
x4_3	1	0.0005	0.0009	-0.0013	0.0023	0.29	0.5908
x4_4	1	0.0006	0.0009	-0.0011	0.0023	0.50	0.4812
x4_5	1	-0.0004	0.0008	-0.0019	0.0011	0.24	0.6258
x5	1	-0.0185	0.0051	-0.0284	-0.0086	13.45	0.0002
x5_1	1	0.0048	0.0054	-0.0057	0.0153	0.80	0.3724
x5_2	1	0.0082	0.0055	-0.0026	0.0190	2.22	0.1360
x5_3	1	-0.0056	0.0054	-0.0162	0.0050	1.07	0.3003
x5_4	1	-0.0044	0.0057	-0.0156	0.0067	0.61	0.4348
x5_5	1	-0.0027	0.0051	-0.0127	0.0073	0.27	0.6013
x6	1	-0.0019	0.0020	-0.0058	0.0020	0.88	0.3477
x6_1	1	0.0015	0.0020	-0.0024	0.0054	0.57	0.4502
x6_2	1	0.0012	0.0019	-0.0026	0.0050	0.39	0.5312
x6_3	1	0.0033	0.0020	-0.0005	0.0072	2.88	0.0898
x6_4	1	0.0052	0.0020	0.0013	0.0092	6.92	0.0085
x6_5	1	-0.0055	0.0019	-0.0093	-0.0017	8.06	0.0045
Dispersion	1	0.0015	0.0005	0.0005	0.0025		

서울시 월별 뇌혈관 질환 사망자의 음이항 회귀 완전모형에서 유의하지 않은 변수를 제거한 모형이 표 3.3에 나타나 있다.

표 3.3 서울시 월별 뇌혈관 질환 사망자에 대한 음이항 회귀 최종모형

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	99	109.9896	1.1110
Scaled Deviance	99	109.9896	1.1110
Pearson Chi-Square	99	110.3797	1.1149
Scaled Pearson X2	99	110.3797	1.1149
Log Likelihood		260411.5998	
Full Log Likelihood		-538.6733	
AIC (smaller is better)		1101.3466	
AICC (smaller is better)		1104.5631	
BIC (smaller is better)		1133.7524	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	6.2423	0.2043	5.8418	6.6428	933.21	<.0001
x1_2	1	-14.5011	5.1236	-24.5432	-4.4591	8.01	0.0047
x1_3	1	11.8414	4.9655	2.1093	21.5735	5.69	0.0171
x2_1	1	-5.1463	2.7313	-10.4995	0.2069	3.55	0.0595
x2_5	1	-10.6478	2.4706	-15.4900	-5.8056	18.58	<.0001
x3_3	1	-6.5425	2.0649	-10.5896	-2.4953	10.04	0.0015
x3_4	1	4.8046	1.9767	0.9304	8.6788	5.91	0.0151
x5	1	-0.0078	0.0019	-0.0116	-0.0041	16.65	<.0001
x6_3	1	0.0030	0.0016	-0.0000	0.0061	3.79	0.0515
x6_4	1	0.0051	0.0016	0.0020	0.0081	10.48	0.0012
x6_5	1	-0.0032	0.0016	-0.0064	-0.0000	3.92	0.0478
Dispersion	1	0.0028	0.0007	0.0015	0.0041		

표 3.3의 추정결과 Dispersion 모수의 추정값이 0.0028로 나타났는데 이는 음이항 분포의 과대산포를 나타내는 모수를 추정한 것으로 $H_0: \gamma = 0$이라는 귀무가설을 검정한다. 이 모수의 추정치는 카이제곱 값을 자유도로 나눈 것의 제곱근 값으로 표 3.3에는 유의확률이 주어져 있지 않지만, 도수자료의 음이항

회귀모형 적합을 위한 또 다른 SAS의 프로시져인 'PROC COUNTREG'를 이용한 결과 유의확률은 $p < 0.0001$로 $H_0: \gamma = 0$이라는 귀무가설을 기각, 과대산포가 있음을 의미한다. 또한 포아송분포는 분산($Var(Y_i|X_i) = \mu_i + \gamma\mu_i^2$)이 평균의 이차식 형태로 나타나는 음이항 분포에서 $\gamma = 0$인 특별한 경우로 위 결과는 서울시 월별 뇌혈관 질환 사망자수가 포아송 분포를 따르지 않음을 의미하기도 한다.

표 3.3의 서울시 월별 뇌혈관 질환 사망자 음이항 회귀 최종모형은 식(3.18)과 같으며 뇌혈관 질환 사망자에 영향을 주는 환경요인들로 아황산가스(SO_2), 오존(O_3), 이산화질소(NO_2), 평균기온, 평균상대습도 등이 있음을 보여주고 있다. 이를 보다 자세히 살펴보면 두달 전, 세달 전 아황산 가스농도($X1$), 한달 전, 다섯달 전 오존농도($X2$), 세달 전, 네달 전의 이산화질소($X3$)농도, 당월의 평균기온($X5$), 세달 전, 네달 전, 다섯달 전의 평균상대습도($X6$)가 서울시 월별 뇌혈관 질환 사망자에 유의한 영향을 나타내는 것으로 보인다.

$$
\begin{aligned}
Log[E(Y_t)] = {} & \underset{}{6.2423} - \underset{(0.2043)}{14.5011}X1_{t-2} + \underset{(5.1236)}{11.8414}X1_{t-3} - \underset{(4.9655)}{5.1463}X2_{t-1} - \underset{(2.7313)}{10.6478}X2_{t-5} \\
& - \underset{(2.4706)}{6.5425}X3_{t-3} + \underset{(2.0649)}{4.8046}X3_{t-4} - \underset{(1.9767)}{0.0078}X5_t \\
& + \underset{(0.0019)}{0.0030}X6_{t-3} + \underset{(0.0016)}{0.0051}X6_{t-4} - \underset{(0.0016)}{0.0032}X6_{t-5} \qquad (3.18)
\end{aligned}
$$

제4장 결론 및 향후 연구과제

본 논문은 서울시 월별 전체 및 65세 이상 뇌혈관 질환 사망자 자료를 이용하여 분해모형 및 자기회귀모형을 적합하였다. 또한 모형의 개선을 위해 자기회귀모형에 대기오염물질(SO_2, O_3, NO_2, PM_{10})과 기상변수(평균기온,평균상대습도)들의 시차변수를 설명변수로 고려한 ADL모형을 적합하여 대기오염물질과 기상변수가 서울시 월별 뇌혈관 질환 사망자에 미치는 영향에 대해 분석하였다. 이에 추가적으로 서울시 월별 뇌혈관 질환 사망자수를 종속변수로 하고 대기오염물질 및 기상변수의 시차변수를 설명변수로 고려한 음이항 회귀모형을 적합하여 분석을 시도하였다.

서울시 월별 뇌혈관 질환 사망자의 분해모형 적합 결과 계절성분은 유의하게 나타났으나 추세성분은 유의하지 않게 나타났으며, 자기회귀모형 적합결과 사망자의 1, 2, 3, 12 시차변수가 유의하게 나타났다. 자기회귀모형에 대기오염물질과 기상변수의 시차변수를 고려한 ADL모형 적합결과 사망자의 1, 2, 3, 12 시차변수와 한달 전과 두달 전 아황산 가스농도가 서울시 월별 뇌혈관 질환 사망자에 유의한 영향을 미치는 것으로 나타났다.

65세 이상 사망자의 경우 전체 사망자의 결과와 달리 분해모형과 자기회귀오차를 가지는 회귀모형에서 추세성분과 계절성분이 모두 유의하게 나타났다. 자기회귀모형 적합결과 사망자의 1, 12 시차변수가 유의하게 나타났으며 ADL모형의 적합결과 전체 사망자의 경우와 유사하게 사망자의 1, 12차 시차변수와 한달 전, 두달 전 아황산 가스농도가 65세 이상 서울시 월별 뇌혈관 질환 사망자에 유의한 영향을 미치는 것으로 나타났다.

서울시 월별 뇌혈관 질환 사망자를 종속변수로 하고 대기오염물질 및 기상변수의 시차변수를 설명변수로 하는 음이항 회귀모형의 최종 적합결과 아황산가스농도($X1_{t-2}, X1_{t-3}$), 오존농도($X2_{t-1}, X2_{t-5}$), 이산화질소($X3_{t-3}, X3_{t-4}$) 농도, 평균기온($X5_t$), 평균상대습도($X6_{t-3}, X6_{t-4}, X6_{t-5}$)가 서울시 월별 뇌혈관 질환 사망자에 유의한 영향을 미치는 것으로 나타났다.

기존연구는 기상요인과 추세변동을 통제한 상태에서의 호흡기 질환 및 순환계 질환 사망자에 대한 단일 대기오염물질의 영향을 분석하였지만 본 논문에서는 좀 더 현실적인 접근을 위해 기상요인 및 추세변동을 통제하지 않고 함께 고려하여 뇌혈관 사망자에 대한 환경요인의 영향을 분석하였다. 시계열 모형에 의한 분석에서는 아황산 가스(SO_2)가 서울시 전체 및 65세 이상 사망자에 유의한 영향을 주는 것으로 나타났다. 이는 아황산가스가 뇌혈관 질환 사망자에 주요한 영향을 미치고 있음을 의미하는 것으로 뇌혈관 질환 사망자에 대한 보건정책수립 시 대기오염물질 중 아황산가스의 관리가 무엇보다도 중요함을 시사하고 있다.

포아송 회귀모형을 이용한 호흡기 질환과 순환계 질환 사망자의 연구결과와 음이항 회귀모형을 이용한 뇌혈관 질환 사망자의 분석결과를 비교하면 미세먼지(PM_{10})는 호흡기 및 순환계 질환 사망자에는 유의한 영향을 주는 반면, 뇌혈관 질환 사망자에는 영향을 주지 않는 것으로 나타났다. 이를 제외하면 사망자에 유의한 영향을 주는 대기오염물질은 대부분 유사하게 나타났다.

시계열모형인 ADL모형과 포아송 회귀모형, 음이항 회귀모형에서 회귀계

수 β는 다음과 같은 해석의 차이가 있다.

시계열모형 중 ADL모형은 연결함수(link function)를 지정해 주지 않은 모형으로 여기서 β는 $\beta = f(Y;X=1) - f(Y;X=0)$로 설명변수가 한 단위 변화 시 Y의 변화량을 의미한다. 그러나 포아송 회귀모형과 음이항 회귀모형은 연결함수를 로그함수로 지정해 준 로그선형모형의 일종으로 식(3.14)와 같이 표현된다. 여기서 β는 $Log[E(Y;X=1)] - Log[E(Y;X=0)] = \beta$를 의미하는 것으로 설명변수가 한 단위 변화 시 $\log[E(Y)]$의 변화를 의미한다. 즉, 설명변수가 한 단위 변화 시 Y가 $\exp(\beta)$만큼 변화함을 의미한다.

본 논문에서 사용된 시계열 분석방법은 기존연구에서 사용한 일반화 가법모형과 비교 시, 향후 사망자 추이를 예측해 볼 수 있다는 장점이 있다. 이러한 예측력은 국가차원의 질병관리 및 정책에 유용하게 사용될 수 있다.

따라서 서울시 월별 전체 뇌혈관 질환 사망자를 자기회귀오차를 가지는 회귀모형, 자기회귀모형, ADL모형에 적합하여 예측을 시도하였다. 표 4.1에 2005년 1월부터 2005년 12월까지의 전체 뇌혈관 질환 사망자의 실제값, 예측값과 예측오차, 그리고 모형 비교를 위한 예측오차에 근거한 RMSE(root mean square error)와 MAPE(mean percentage error)가 주어져 있다.

표 4.1 모형에 의한 예측 값, 예측오차와 모형비교

날짜	Y_t	자기회귀오차를 가지는 회귀모형		AR모형		ADL모형	
		예측 값	예측오차	예측 값	예측오차	예측 값	예측오차
05.01	473	510.14	-37.14	505.81	-32.81	511.24	-38.24
05.02	376	441.36	-65.36	460.91	-84.91	460.17	-84.17
05.03	455	497.96	-42.96	484.47	-29.47	447.27	7.73
05.04	496	468.52	27.48	424.87	71.13	417.06	78.94
05.05	483	463.71	19.29	478.34	4.66	493.46	-10.46
05.06	403	419.10	-16.10	452.43	-49.43	470.74	-67.74
05.07	416	409.66	6.34	461.38	-45.38	436.51	-20.51
05.08	411	405.59	5.41	457.49	-46.49	440.60	-29.60
05.09	387	423.55	-36.55	447.31	-60.31	436.93	-49.93
05.10	473	498.03	-25.03	468.15	4.85	453.36	19.64
05.11	441	494.45	-53.45	486.73	-45.73	487.84	-46.84
05.12	442	516.79	-74.79	477.57	-35.57	485.13	-43.13
AIC		1181.07		1203.12		1055.18	
SBC		1220.09		1217.06		1073.82	
RMSE		40.17		48.23		48.10	
MAPE		7.90		10.07		9.72	

표 4.1의 결과 모형 선택의 기준이 되는 AIC통계량을 기준으로 보면 ADL모형이 가장 우수한 것으로 보이나, 예측의 관점에서 각 모형의 예측오차에 근거한 RMSE와 MAPE를 비교해 보면 자기회귀오차를 가지는 회귀모형이 가장 우수한 것으로 보인다. 여기서 서울시 월별 뇌혈관 질환 사망자의 최종모형으로 자기회귀오차를 가지는 회귀모형과 ADL모형 중 한 모형을 선택하는 일은 의미가 없다고 보여진다. 연구의 목적에 따라 향후 사망자를 예측하여 정책수립을 목적으로 하는 연구의 경우 예측력이 우수한 자기회귀오차를 가지는 회귀모형이 최종모형으로 선택되어질 것이고 사망자에 영향을 주는 환경요인을 찾아 이를 관리하는 것을 목적으로 하는 연구의 경우 ADL모형이 최종모형으로 선택되어질 것이다.

본 논문의 경우 ADL모형에서 설명변수 차수를 표 2.8에 주어진 것과 같이 상관분석을 통해 모형의 효율성과 편의성을 고려하여 결정하였다. 이렇게 결정된 시차변수를 가지고 그랜져 인과성 검정을 실시하였는데 그랜져 인과성 검정결과는 설명변수의 시차길이에 따라 결과가 달라지므로 시차길이 선택은 최종모형 선택에 매우 중요한 역할을 한다. 따라서 좀 더 정확한 연구·분석을 위해 향후 연구에서는 시차길이 선정에 일반적으로 사용되는 FPE(Final prediction error)와 BEC(Bayesian estimation criterion)통계량에 근거한 시차길이를 선택하여 모형을 적합 하는 것이 필요할 것이라 생각된다(Daniel and Dalls, 1985).

본 논문에서 제시한 뇌혈관 질환은 한국인의 3대 주요 사인으로 계속해서 증가하고 있는 추세이다. 따라서 본 논문을 바탕으로 뇌혈관 질환에 영향을 미치는 대기오염물질을 배출량을 체계적으로 관리하고, 조기 건강검진과 경고시스템을 구축하여 뇌혈관 질환의 발생을 조기에 예방하는 일이 무엇보다도 중요하다고 생각되어 진다.

마지막으로 향후 연구에서는 음이 아닌 정수 값인 사망자수를 대상으로로 'Integer valued time series model'를 이용하여 개선된 예측결과를 얻는 연구·분석이 수행되어져야 한다고 생각한다.

ABSTRACT

Time Series Analysis on the Effect of Environmental Factors on the death of Cerebrovascular Diseases

Heewon Park
Department of Statistics
The Graduate School
Sungshin Women's University

Serious problems caused by environmental pollutants are due to rapid industrialization and the cityward tendency of the population. Especially air pollution has caused severe health problems.

The purpose of this study is to analyze the effects of environmental changes on the deaths of cerebrovascular diseases using mortality data, air pollutants and climate data from January 1995 to December 2004.

Statistical methods such as 'regression model', 'autoregressive regression model with errors', 'autoregressive model', 'autoregressive distributed lag model' and 'negative binomial regression model' are used to evaluate the association between environmental factors and the mortality.

As a result, SO_2 in autoregressive distributed lag model and SO_2, O_3,

NO_2, *Temperature* and *Humidity* in negative binomial model affect the monthly death counts of cerebrovascular disease.

Based on AIC and SBC, 'autoregressive distributed lag model' is better than other models for the death counts of cerebrovascular disease. On the other hand 'autoregressive regression model with errors model' is better in the viewpoint of forecasting based on RMSE and MAPE.

참 고 문 헌

[1] 권민경 (2003). 포아송 회귀모형과 음이항 회귀모형의 비교연구, 고려대학교, 통계학과 석사학위 논문, 서울.

[2] 권호장,조수헌(1999). 서울시 대기오염과 일별 사망자 수의 관련성에 대한 시계열적 연구, 예방의학학회지, 32, 191-199.

[3] 김윤신,최원우,김무채(1998). 공단지역 대기오염과 일별사망자수와의 연관성에 대한 연구, 한국보건통계학회지, 23, 124-236.

[4] 심정욱,손여숙,백장선.(2003). 수리통계학, 자유아카데미, 서울.

[5] 양희은(2004). 일반화 가법모형을 이용한 대기오염과 사망자수와의 연관성, 덕성여자대학교, 정보통계학과 석사학위 논문, 서울.

[6] 이승욱(2004). 부산시 대기오염에 따른 사망양상, 인제대학교, 의학과 박사학위 논문, 김해.

[7] 이종협(2007). 시계열분석과 응용, 자유아카데미, 서울.

[8] 전명식(2007). 수리통계학, 자유아카데미, 서울.

[9] 통계청(1994-2004). 사망원인통계.

[10] 통계청(2000-2004). 주민등록 인구통계.

[11] 통계청(2004). 한국표준질병·사인분류 제2권지침서, 8-130.

[12] Akaike, H.(1973). Information theory and an extension of the maximum likelihood principal. *Proc. 2nd International Symposium Information Theory.* 267-281. Akademiai Kiado, Budapest.

[13] Akaike, H.(1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving avsrage processes, *Annals of the Institute of Statistical Mathematics.* 36. 363-387.

[14] Cameron, A.C. and Trivedi, P.K.(1998). *Regression analysis of count data.* Cambridge University, New York.

[15] Conover, W.J.(1980). *Practical Nonparametric Statistics.* John wiley & Sons, New York.

[16] Daniel, L.T. and Dalls, S.B.(1985). Lag-Length Selection and Tests of Granger Causaluity Between Money and Income. *Journal of Money, Credit and Banking,* 17, 164-178.

[17] Granger, C.W.J(1969). Investigating Causal Relations by Econometric Models and Cross-Spectral Methods, *Econometrica,* 37, 124-438.

[18] Ismail, N. and Jemain, A.A.(2007). Handling Overdispersion with Negative binomial and Generalized Poisson Regression Models. *Casualty Actuarial Society Forum,* 103-158.

[19] Paul, D.A.(1999). *Logistic Regression Using the SAS System : Theory and Application*, SAS Institute.

[20] SAS Institute Inc(2008). *SAS/STAT User's Guide 9.2*.

[21] Stock, J.H and Watson, M.W.(2002). *Introduction to Econometrics,* Addison Wesley, New York.

[22] Wei, W.W.S.(2006). *Time Series Analysis* : *Univariate and Multivariate Methods,* 2nd Ed. Addison Wesley, New York.