이 종 협 교수지도

석사학위청구논문

한국인의 주요 사망원인에 대한 시계열 자료 분석

2007

성신여자대학교 대학원

통 계 학 과

김 영 은

한국인의 주요 사망원인에 대한 시계열 자료 분석

이 종 협 교수지도

이 논문을 석사학위논문으로 제출함

2006년 11월

성신여자대학교 대학원

통 계 학 과

김 영 은

인 준 서

김영은의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

우리나라의 경우 세계적으로 유례없는 빠른 경제성장과 산업화로 인하여 사인구조가 급격히 변화하고 있으며, 고령 사회로의 진입 초기 단계인 현 시점에서 사인구조를 시계열적으로 분석하여 추이를 예측하는 것은 매우 중요하다.

본 논문에서는 1995년부터 2004년까지 사망원인 통계자료를 이용하여 주요 사망원인에 대한 전반적인 경향을 빈도분석 및 교차분석, 수량화방법Ⅱ를 이용하여 탐색적으로 살펴본 후 사인 순위 선정을 위한 56항목에 해당하는 개별 사인에 대하여 ARIMA모형을 적합 시켜 보고 향후 추이를 예측한다.

또한 우리나라 사망원인 1위인 악성신생물에 대해서 각 장기별로 ARIMA 모형에 의한 예측을 통해 향후 사망추이를 전망한다. 최근 높은 증가율을 보이고 있는 대장암에 대하여 ARIMA 모형, 평활법, 분해법을 이용하여 예측력을 비교한 후 최적모형을 선택하고 그 전개추이를 고찰한다.

목 차

제1장 서 론

사망원인 통계는 국가와 사회 집단의 건강 및 보건 상태를 나타내 주며, 사회 문화적 지표가 된다. 또한 국민 보건 및 의료 분야에서의 정책 수립과 연구에 중요한 자료로 활용되기도 한다. 일반적으로 사망수준이나 사망구조 또는 사망원인은 한 시대의 사회, 경제 구조나 환경적 여건에 따라서 달리 한다.

UN의 보고에 의하면, 전통사회에서의 사인 양상은 전염성 또는 호흡기계 질환에 의해 높은 사망률을 나타내고 있는 반면 오늘날은 예방 및 치료의학의 발달로 인위적으로 통제할 수 있는 질병을 해결함으로써 중년기 이후의 비전염성질환인 순환기계 질환, 손상 및 중독 등에 의한 사망률이 증가하고 있다고 보고하고 있다.

우리나라의 경우 세계적으로 유례없는 빠른 경제성장과 산업화로 인하여 사인 구조가 급격히 변화하고 있으며, 고령화 사회로의 진입 초기 단계인 현 시점에서 사인 구조를 시계열적으로 분석하여 추이를 예측하는 것이 매우 중요하다. 따라서 본 논문에서는 우리나라 사망 원시자료를 토대로 주요 사망원인에 대한 기초분석을 실시하여 사망원인의 특성을 살펴보고, 시계열 분석을 통하여 적합모형을 제시한 후 향후 추이를 예측하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 사망원인 자료에 대한 기초분석으로 개괄적인 특성을 알아본다. 제3장에서는 사인 순위 선정을 위한 56항목에 해당하는 개별 사인과 우리나라 사망원인 1위인 악성신생물(암)에 대해서 ARIMA모형을 적합 시켜 본다. 또한 악성신생물 중에서 대장암에 대하여 ARIMA모형, 평활법, 분해모형을 이용한 예측을 실시한 후 예측력을 비교해 본다. 마지막으로 4장에서는 본 연구의 결론을 맺는다.

제2장 기초분석

2.1 분석 데이터의 특징

우리나라의 사망 원인통계는 국민의 정확한 사망원인 구조를 파악하여 국민복지 및 보건의료 정책수립을 위한 기초자료로 활용하기 위하여 1년 동안 국민이 제출한 사망신고서를 기초로 사망자의 사망원인을 세계보건기구(WHO)의 국제질병분류(International Classification of Diseases; ICD-10) 수정내용을 반영한 한국 표준 질병 사인분류(Korean Classification of Diseases; KCD) 체계에 의해 의거 분류·집계 된다.

사망자 발생시 작성된 사망신고서 기준으로 데이터는 수집되어지는데 이에 적힌 항목들은 표2.1과 같다.

표2.1 사망신고서에 기재되어 있는 항목

항목	하위항목
1. 일반 사항	· 사망 연도 · 사망 월 · 사망 일
	· 성별(남/여)
	· 연령
	· 결혼 상태(미혼/유배우/이혼/사별/미상)
	· 직업
	· 교육정도
2. 사망지역과 거주지	· 신고지
	· 신고 일자
	· 주소지
	· 사망 장소
3. 사망원인	· 사망분류기준(KCD) 정보

사망원인 정보인 KCD는 의무 기록자료 및 사망원인 통계조사 등 질병이환 및 사망 자료를 그 성질의 유사성에 따라 전염성 질환, 체질적 또는 전신적 질환, 부위에 따른 국소질환, 발육질환, 손상의 항목으로 체계적으로 유형화하여 모든 형태의 보건 및 인구동태 기록에 기재되어 있는 질병 및 기타 보건문제를 분류하는데 이용하기 위하여 설정되었다. WHO에서 제 10차 국제질병분류의 수정판(ICD-10)을 작성하여 각 회원국에 적용하도록 권고하여 2003년부터 제 4차 한국 표준 질병 사인분류(KCD-4)를 사용하고 있다. KCD는 대분류 21개, 중분류 261개, 소분류 2,036개, 세분류 12,171개의 분류로 이루어져 있다. 사망진단서에 기재되어야 할 사인(cause of death)은 '사망을 초래하거나 원인이 되는 모든 질병, 병태 또는 손상과 이러한 손상을 유발시킨 사고나 폭력상황'이라고 정의하고 있으며, 일차 제표를 위하여 선정되는 원사인(Underlying cause of death)은 '직접적으로 사망에 이르게 한 일련의 사건들을 야기한 질병 또는 손상', '치명상을 일으킨 사고나 폭력상황'으로 정의하고 있다. 그러나 근원적인 선행 사인이 XIX 장에 분류되는 '손상, 중독 및 외인에 의한 특정 기타 결과' 인 경우 그러한 병태를 유발시킨 상황이 제표용 원사인으로 선정되어야 한다. KCD의 분류구조는 표2.2와 같다.

사망원인 통계는 19개의 장(章), 103항목(WHO에서 유의하여 볼 필요성이 있는 사인으로 권고한 항목), 56항목(사인순위 선정을 위한 항목), 236항목(우리나라에서 많이 발생하는 사인을 위주로 한 항목)별로 구분하여 제표하고 있다.

우리나라에서 발표되는 사망원인 통계 자료는 56항목에 의한 것이 대부분인데 이는 통계청에서 사인순위를 발표할 때 56항목을 기준으로 순위를 부여하여 발표하기 때문이다. 향후 발표될 사망원인 통계와 쉽게 비교 가능하게 하기 위하여 본 논문에서는 56항목에 해당하는 10년간 사망자수(1995년

부터 2004년 사이에 사망신고서를 기초로 집계된 자료)를 분석 자료로 사용한다.

56항목에는 전신을 침해한 질환군이 25개, 정신병적 질환군이 3개, 인체 해부학적 계통별 질환군이 15개, 분만·기형·신생아 질환군이 5개, 기타병태가 1개, 기타분류가 7개가 포함되어 있다.

표2.2 한국 표준 질병·사인의 분류(KCD) 구조

전신을 침해한 질환군	Ⅰ 특정 감염성 및 기생충성 질환
	Ⅱ 신생물
정신병적 질환군	Ⅲ 혈액 및 조혈기관의 질환과 면역기전을 침범하는 특정 장애
	Ⅳ 내분비, 영양 및 대사 질환
인체 해부학적 계통별 질환군	Ⅴ 정신 및 행동 장애
	Ⅵ 신경계통의 질환
	Ⅶ 눈 및 눈 부속기의 질환
	Ⅷ 귀 및 꼭지돌기의 질환
	Ⅸ 순환기계통의 질환
	Ⅹ 호흡기계통의 질환
	Ⅺ 소화기계통의 질환
	Ⅻ 피부 및 피부 밑 조직의 질환
	ⅩⅢ 근육골격계통 및 결합조직의 질환
	ⅩⅣ 비뇨생식기계통의 질환
분만·기형·신생아 질환	ⅩⅤ 임신, 출산 및 산후기
	ⅩⅥ 출생전후기에 기원한 특정 병태
	ⅩⅦ 선천 기형, 변형 및 염색체 이상
기타 병태	ⅩⅧ 달리 분류되지 않은 증상, 징후와 임상 및 검사의 이상 소견
	ⅩⅨ 손상, 중독 및 외인에 의한 특정 기타 결과
기타 분류	ⅩⅩ 질병이환 및 사망의 외인
	ⅩⅩⅠ 건강상태 및 보건서비스 접촉에 영향을 주는 요인

2.2 사망원인 자료에 대한 빈도분석 및 교차분석

현대의 한국 사회는 다변화되고 복잡해지는 사회 구조 속에서 정신적인 스트레스와 음주, 흡연의 증가뿐만 아니라 생활습관과 식습관이 점차 서구적으로 변모하는 양상을 보인다. 빈도분석과 교차분석을 통해 이러한 인구사회적 변화에 따른 사망원인의 특성을 파악해 보자.

2.2.1 빈도분석

1995년 1월부터 2004년 12월까지 국민이 제출한 사망신고서를 기초로 사망자의 사망원인을 한국 표준 질병 사인분류 체계에 의해 집계한 통계 자료를 분석한 결과 총 2,435,678명이 사망하였으며 이중 56항목에 분류되어 지는 사망자 수는 1,838,575명으로 75.49%이다.

표2.3은 10년 동안 사망자수에 순위를 부여하여 재 정렬한 후 1995년과 2004년의 각 사인별 빈도를 정리해 놓은 표이다. 1위는 악성신생물(암)이며 전체 사망자 가운데 23.2%(565,985명)를 차지하고 있고, 2위는 뇌혈관질환으로 14.40%(350,752명)을 차지하고 있다. 이 두 가지 질환이 전체 사망자의 약 38%를 차지하고 있어 우리나라의 대표적인 질환이라 불리어져도 과언이 아니다. 또한 상위 10순위에 의한 사망자가 전체 사망자 2,435,687명중 68.4%(1,665,697명)를 차지하고 있다.

악성신생물과 당뇨병, 자살은 95년에 비하여 크게 증가하였으며, 운수사고와 고혈압성 질환은 감소하였다. 특히 후진국병으로 잘 알고 있는 호흡기 결핵이 상위 10위안에 들고, 10년 전에 비해 크게 줄지 않은 것은 경제의

비약적인 발전에 비하여 의료 및 보건 분야의 취약함이 있음을 증명해 준다. 우리나라는 결핵으로 인한 사망이 OECD국가 가운데 1위인 것이 이를 뒷받침 해준다. 또한 자살로 인한 사망자가 급증하였는데 이는 우리나라가 지난 10년 동안 외환위기를 겪으면서 가족의 붕괴나 경제적인 곤란 등의 사회 구조, 가치관의 변화와 연관이 있을 것으로 생각된다. 반면에 교통사고로 인한 사망자수는 10년 전에 비해 절반이상 감소하였다.

특히 영아사망에 관한 사인(출생전후기질환, 선천기형)은 1995년과 2004년에 큰 폭으로 변화하였는데 사망특성상 신고 되지 않는 경우가 많아 1999년부터 외부 행정기관 자료를 이용하여 보완하였다. 분석에는 1999년 이후 자료를 이용하도록 하겠다. 그림2.1은 상위 10순위에 대한 연도별 사망자수의 그래프이다.

표2.3 사인순위 선정을 위한 56항목

순위	KCD code	사망원인	빈도(명)	
			1995년	2004년
1	C00-C97	악성신생물	50,107	64,731
2	I60-I69	뇌혈관 질환	36,061	34,091
3	I20-I51	심장 질환	16,682	17,915
4	V01-V99	운수사고	17,497	8,333
5	K70-K76	간 질환	13,323	9,272
6	E10-E14	당뇨병	7,789	11,768
7	J40-J47	만성 하기도 질환	6,763	8,378
8	X60-X84	자살	4,841	11,523
9	I10-I13	고혈압성 질환	8,276	5,036
10	A15-A16	호흡기 결핵	3,739	2,780
11	J12-J18	폐렴	1,909	3,512
12	W00-W19	추락	2,391	3,346
13	W65-W74	익사	1,774	964
14	F10-F19	정신활성물질 사용에 의한 정신 및 행동장애	1,193	1,012
15	A40-A41	패혈증	669	936

순위	KCD code	사망원인	빈도(명)	
			1995년	2004년
16	P00-P96	출생전후기질환	248	1,197
17	Q00-Q99	선천 기형	992	734
18	X85-Y09	타살	818	844
19	X40-X49	중독사고	1,330	256
20	K25-K27	위 및 십이지장 궤양	817	464
21	X00-X09	화재사고	849	372
22	B15-B19	바이러스 간염	197	801
23	G30	알쯔하이머병	10	1,241
24	D50-D64	빈혈	275	258
25	I70	죽상경화증(동맥경화증)	737	181
26	Re. A00- B99	나머지 특정 감염성 및 기생충성 질환	141	319
27	N00-N15	사구체질환 및 세뇨관-사이질성 질환	288	149
28	I00-I09	급성 류마티스열 및 만성 류마티스 심장 질환	206	291
29	A17-A19	기타 결핵	190	168
30	E40-E46	영양실조	181	52
31	J20-J22	기타 급성 하기도 감염	23	73
32	G00,G03	수막염	194	57
33	A09	감염성 기원으로 추정되는 설사 및 위장염	161	60
34	R95	영아 급사 증후군	122	70
35	J10-J11	인플루엔자	156	10
36	A01-A08	기타 창자 감염성 질환	68	30
37	O10-O92	기타 직접산과적 사망	2	0
38	B20-B24	인체 면역결핍 바이러스병	13	72
39	A90-A94, A96-A99	기타 절지동물 매개의 바이러스열 및 바이러스출혈열	44	12
40	A33-A35	파상풍	12	5
41	A50-A64	주로 성행위로 전파되는 감염	5	2
42	B05	홍역	14	0
43	B50-B54	말라리아	5	4
44	A39	수막알균감염	5	1
45	O00-O07	유산된 임신	5	0
46	O98-O99	간접산과적 사망	2	0
47	A80	급성 회색질척수염	1	0
48	A82	광견병	0	1
49	B65	주혈흡충증	1	0
50	A00	콜레라	0	0

순위	KCD code	사망원인	빈도(명)	
			1995년	2004년
51	A36	디프테리아	0	0
52	A95	황열	0	0
53	B55	리슈만편모충증	0	1
54	A20	페스트	0	0
55	A37	백일해	0	0
56	B56-B57	파동편모충증	0	0
합 계			**181,198**	**191,377**

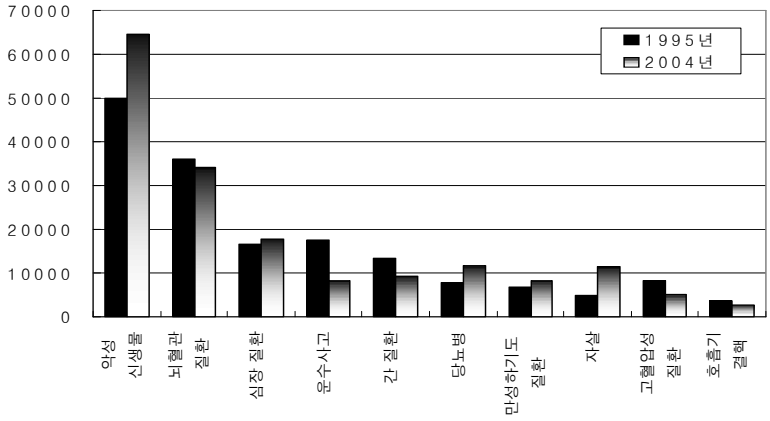


그림2.1 상위10위까지의 연도별 비교

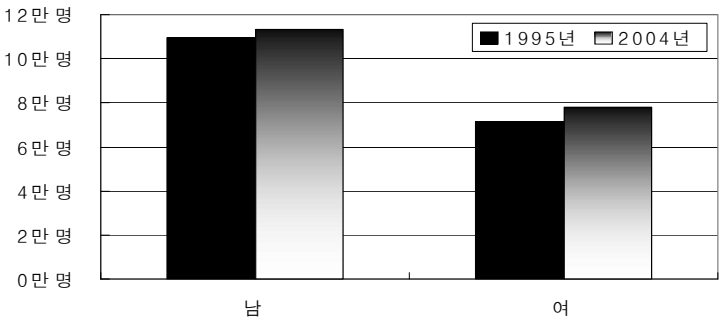
표2.4는 1995년과 2004년의 성, 연령, 학력, 결혼 상태, 교육 수준별 사망자수이며 그림2.2는 이에 대한 그래프이다.

전체 사망자는 남성의 빈도가 여성보다 많으며, 1995년보다 남녀 모두 사망자수가 소폭 증가하였다. 연령별로는 70대의 사망자수가 가장 많으며 2004년에는 1995년에 비해 60대 이상 연령의 사망자수가 더 많았다. 이는 인구 구조 고령화에 기인한 것으로 생각된다. 교육수준별로는 무학 또는 초등학교 졸업의 사망자수가 많으며 고졸이상의 사망자수가 1995년에 비해 증

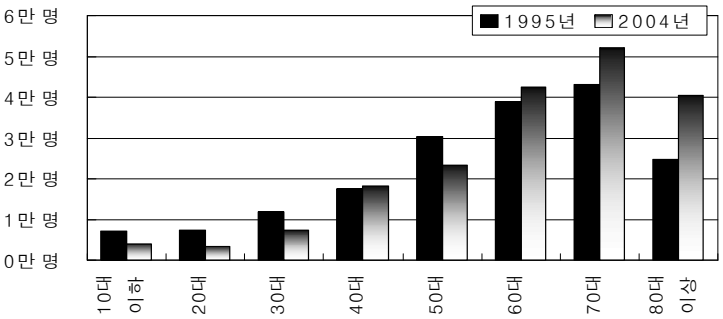
가하였다. 가장 많은 사망자수를 보이는 70~80대 고령자의 교육수준이 낮아 이러한 결과가 나온 것으로 생각된다. 결혼 상태별로는 배우자가 있거나 사별인 사망자가 많으며 1995년과 2004년에 빈도의 차이는 크지 않은 것으로 판단된다.

표2.4 성, 연령, 학력, 결혼상태, 교육수준별 사망자수

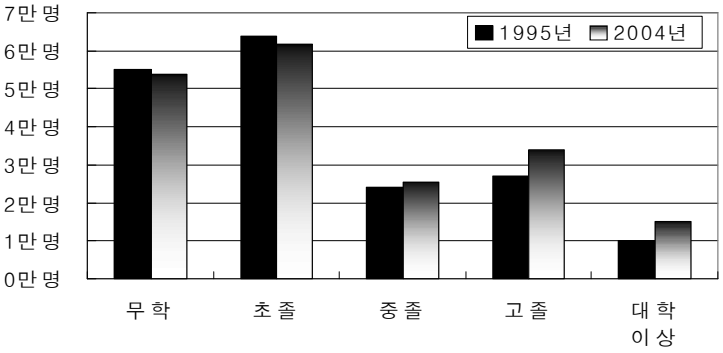
범 주		빈 도 (명)		
		1995년	2004년	전체
성별	남	109,800	113,408	1,358,356
	여	71,398	77,969	1,077,331
연령	10대 이하	7,254	4,079	72,675
	20대	7,486	3,305	63,156
	30대	11,983	7,360	115,293
	40대	17,492	18,152	204,159
	50대	30,295	23,440	296,178
	60대	38,823	42,434	467,808
	70대	43,164	52,109	604,442
	80대 이상	24,701	40,494	611,857
	미상	0	4	119
교육 수준	무학	55,100	53,772	879,742
	초등학교 졸업	63,821	61,759	765,560
	중학교 졸업	24,248	25,512	288,248
	고등학교 졸업	27,065	33,867	347,100
	대학 이상	10,108	15,143	144,423
	미상	856	1,324	10,614
결혼 상태	미혼	20,091	15,354	218,005
	배우자 있음(유배우)	99,264	102,595	1,206,357
	이혼	4,457	9,302	78,816
	사별	54,815	64,158	907,364
	미상	2,571	2,668	25,145


12만 명
10만 명
8만 명
6만 명
4만 명
2만 명
0만 명
1995년
2004년
남
여

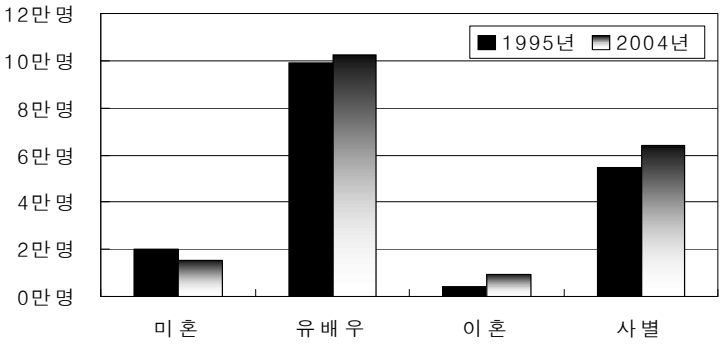
<성>


6만 명
5만 명
4만 명
3만 명
2만 명
1만 명
0만 명
1995년
2004년
10대 이하
20대
30대
40대
50대
60대
70대
80대 이상

<연령>


7만 명
6만 명
5만 명
4만 명
3만 명
2만 명
1만 명
0만 명
1995년
2004년
무학
초졸
중졸
고졸
대학 이상

<교육 수준>



<결혼 상태>

그림2.2 성, 연령, 학력, 결혼 상태, 교육 수준별 분포

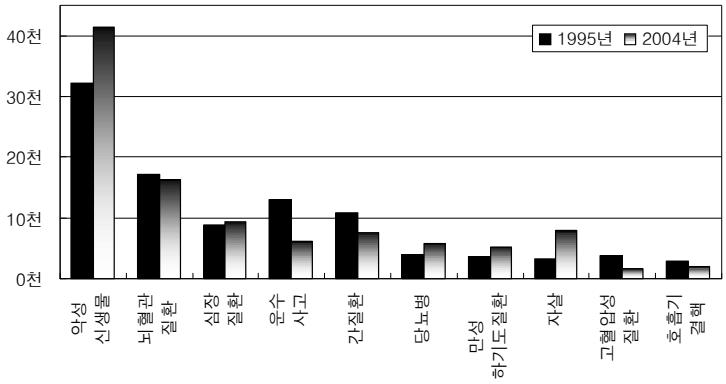
2.2.2 교차분석

이제 각 사인에 대하여 성, 연령, 결혼여부, 교육정도에 따른 사망자수를 비교해보자. 표2.5는 성별에 따른 주요 10대 사망원인에 대한 빈도 및 순위표이며, 그림2.3은 그에 대한 그래프이다. 악성신생물, 심장질환, 운수사고, 간질환, 당뇨병, 자살, 호흡기 결핵은 남성 사망자가 많으며, 뇌혈관질환과 고혈압성질환은 여성 사망자가 더 많다. 성별에 따른 순위 변화를 살펴보면 간질환과 고혈압성 질환이 남녀별로 뚜렷한 순위변화가 관찰되며, 특히 고혈압성 질환의 경우 남성 사망자수의 하락폭이 여성 사망자수의 하락폭보다 작았다.

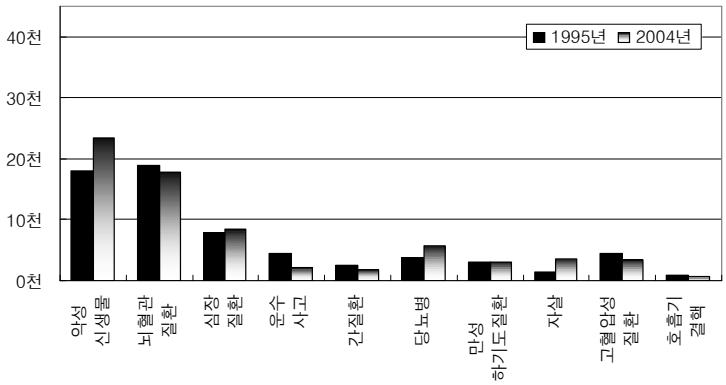
성별과 사망원인의 독립성 검정결과 피어슨 카이제곱 통계량이 1995년과 2004년에서 모두 유의하게 나타났다. 이는 성별에 따라 각 사인별로 유의한 차이를 보이고 있음을 의미한다.

표2.5 성별에 따른 10대 사인에 대한 빈도와 순위

	1995년		2004년	
	남성(순위)	여성(순위)	남성(순위)	여성(순위)
악성신생물	32,109(1)	17,998(2)	41,312(1)	23,419(1)
뇌혈관질환	17,178(2)	18,883(1)	16,219(2)	17,872(2)
심장질환	8,837(5)	7,845(3)	9,443(3)	8,472(3)
운수사고	12,981(3)	4,516(5)	6,125(6)	2,208(8)
간질환	10,857(4)	2,466(8)	7,549(5)	1,723(9)
당뇨병	3,957(6)	3,832(6)	5,858(7)	5,710(4)
만성하기도질환	3,669(8)	3,094(7)	5,228(8)	3,150(7)
자살	3,320(9)	1,521(9)	7,903(4)	3,620(5)
고혈압성질환	3,726(7)	4,550(4)	1,684(12)	3,352(6)
호흡기결핵	2,831(10)	908(10)	1,975(10)	802(12)
χ^2	10178 (p<.0001)		9224 (p<.0001)	



<남성>



<여성>

그림2.3 10대 주요사인에 대한 남녀별 비교

남성의 경우 과음과 흡연, 불규칙한 생활습관, 스트레스 등의 이유로 운수사고나 간질환, 자살의 빈도가 여성보다 높다고 생각되어 지며, 여성의 경우 비만, 임신과 출산, 폐경 등의 이유로 고혈압성 질환이 높은 것으로 생각되어 진다. 그러나 남성의 사망빈도가 높았다고 해서 남성에게 더 빈번한 사인이라는 것을 의미하지는 않는다. 이를 통합적으로 고려하기 위해서는 성비에 따른 가중치를 두는 간접 또는 직접표준화 사망률을 고려하는 것이 바람직하다.

표2.6는 각 연령그룹에서 가장 많이 발생한 사인을 정리한 표이다. 20대 이하 연령대에서는 운수사고 및 자살, 추락 등 사고사로 인한 사망자가 많았고, 30대 이상 연령대에서는 악성신생물과 뇌혈관 질환으로 인한 사망자가 가장 많았다.

사망원인이 연령에 따라 차이를 보이고 있는지 알아보기 위하여 순서형 범주에 사용하는 코크란-맨텔-헨첼통계량(CMH; M^2)을 이용하여 검정해본 결과 1995년과 2004년에서 모두 유의하므로 연령대가 높아질수록 사망자수

가 늘어가는 경향을 보이고 있음을 의미한다.

표2.6 연령별 사인 순위

		1995년							
		10대 이하	20대	30대	40대	50대	60대	70대	80대 이상
1위	사인	운수사고	운수사고	운수사고	악성신생물	악성신생물	악성신생물	뇌혈관질환	뇌혈관질환
	빈도	2591(35)	3362(45)	3041(25)	5061(29)	11314(37)	14392(37)	12810(30)	8093(33)
2위	사인	선천기형	자살	악성신생물	간질환	뇌혈관질환	뇌혈관질환	악성신생물	심장질환
	빈도	884(12)	1105(15)	2340(20)	2978(17)	4311(14)	8268(21)	11972(28)	3737(15)
3위	사인	악성신생물	악성신생물	간질환	운수사고	간질환	심장질환	심장질환	악성신생물
	빈도	712(10)	814(11)	1416(12)	2484(14)	3982(13)	3368(9)	4380(10)	3502(14)
4위	사인	익사	심장질환	자살	뇌혈관질환	운수사고	간질환	고혈압성질환	고혈압성질환
	빈도	633(9)	370(5)	1080(9)	1661(10)	2481(8)	2665(7)	2665(6)	2506(10)
5위	사인	자살	익사	심장질환	심장질환	심장질환	당뇨병	당뇨병	만성하기도질환
	빈도	347(5)	328(4)	903(8)	1369(8)	2248(7)	2187(6)	2551(6)	2357(10)
6위	사인	심장질환	추락	뇌혈관질환	자살	당뇨병	고혈압성질환	만성하기도질환	당뇨병
	빈도	207(4)	231(3)	629(5)	775(4)	1284(4)	1819(5)	2551(6)	1001(4)
7위	사인	출생전후기질환	타살	호흡기결핵	당뇨병	고혈압성질환	운수사고	간질환	간질환
	빈도	248(3)	194(3)	351(3)	496(3)	857(3)	1803(5)	1500(3)	604(2)
8위	사인	추락	뇌혈관질환	추락	호흡기결핵	호흡기결핵	만성하기도질환	운수사고	폐렴
	빈도	225(3)	182(2)	351(3)	487(3)	680(2)	1128(3)	1274(3)	544(2)
9위	사인	화재사고	중독사고	익사	행동장애	자살	호흡기결핵	호흡기결핵	운수사고
	빈도	174(2)	170(2)	252(2)	346(2)	677(2)	776(2)	865(2)	461(2
10위	사인	폐렴	호흡기결핵	중독사고	추락	만성하기도질환	자살	폐렴	호흡기결핵
	빈도	162(2)	133(2)	236(2)	327(2)	440(1)	435(1)	542(1)	411(2)
M^2		10723 (p<.0001)							

		2004년							
		10대 이하	20대	30대	40대	50대	60대	70대	80대 이상
1위	사인	출생전후기질환	자살	자살	악성신생물	악성신생물	악성신생물	악성신생물	뇌혈관질환
	빈도	1197(29)	1088(33)	1829(15)	5915(33)	9946(42)	18739(45)	18370(35)	10611(26)
2위	사인	운수사고	운수사고	악성신생물	간질환	뇌혈관질환	뇌혈관질환	뇌혈관질환	악성신생물
	빈도	651(16)	875(27)	1808(25)	2514(14)	2500(11)	6869(16)	12064(23)	9030(22)
3위	사인	선천기형	악성신생물	운수사고	자살	간질환	심장질환	심장질환	심장질환
	빈도	606(15)	472(14)	1003(14)	2419(13)	2438(10)	3447(8)	5178(10)	5394(13)
4위	사인	악성신생물	심장질환	간질환	뇌혈관질환	자살	당뇨병	당뇨병	만성하기도질환
	빈도	451(11)	180(6)	550(8)	1526(8)	1822(8)	3220(8)	4185(8)	3551(9)
5위	사인	익사	익사	심장질환	운수사고	심장질환	간질환	만성하기도질환	당뇨병
	빈도	248(6)	126(4)	493(7)	1484(8)	1803(8)	2121(5)	3116(6)	2327(6)
6위	사인	자살	추락	뇌혈관질환	심장질환	당뇨병	자살	고혈압성질환	고혈압성질환
	빈도	247(6)	94(3)	387(5)	1326(7)	1230(5)	1875(4)	1581(3)	2318(6)
7위	사인	타살	타살	추락	당뇨병	운수사고	운수사고	자살	폐렴
	빈도	100(3)	85(3)	235(3)	625(3)	1170(5)	1575(4)	1446(3)	1879(5)
8위	사인	추락	뇌혈관질환	타살	추락	추락	만성하기도질환	간질환	추락
	빈도	94(2)	75(2)	166(2)	414(2)	407(2)	1238(3)	1102(2)	992(2)
9위	사인	화재사고	호흡기결핵	호흡기결핵	행동질환	호흡기결핵	고혈압성질환	운수사고	자살
	빈도	93(2)	59(2)	154(2)	344(2)	323(1)	20(2)	1082(2)	797(2)
10위	사인	영아급사증후군	간질환	당뇨병	호흡기결핵	만성하기도질환	추락	폐렴	알쯔하이머병
	빈도	70(2)	29(1)	153(2)	322(2)	288(1)	496(1)	991(2)	750(2)
M^2		16997 (p<.0001)							

※ ( ) : 각 연령 범주에서 해당사인이 차지하고 있는 비율(%)

표2.7는 교육수준에 대하여 상위 10개의 사망원인을 정리한 표이다. 교육수준은 순서형 변수 이므로 사망원인별로 유의한 차이가 있는지 알아보기 위하여 코크란-맨텔-헨첼 통계량(CMH; M^2)을 이용한다. 검정 결과 1995년과 2004년에서 모두 통계적으로 유의하였다. 즉 교육수준에 따라 사인별로 유의한 차이가 존재함을 의미한다.

교육 수준이 높아질수록 운수사고나 자살로 인한 사망자 비율이 증가하였

다.

표2.7 교육수준별 사인 순위

		1995년				
		무학	초등학교 졸업	중학교 졸업	고등학교 졸업	대학 이상
1위	사인	뇌혈관질환	악성신생물	악성신생물	악성신생물	악성신생물
	빈도	15406(28)	19581(31)	7347(30)	7782(29)	3465(34)
2위	사인	악성신생물	뇌혈관질환	뇌혈관질환	운수사고	운수사고
	빈도	11692(21)	12648(20)	3501(14)	5604(21)	1667(17)
3위	사인	심장질환	간질환	운수사고	뇌혈관질환	뇌혈관질환
	빈도	6047(11)	5756(9)	2982(12)	3067(11)	1271(13)
4위	사인	고혈압성질환	심장질환	간질환	간질환	심장질환
	빈도	4334(8)	5643(9)	2572(11)	2047(8)	946(9)
5위	사인	만성하기도질환	운수사고	심장질환	심장질환	간질환
	빈도	3423(6)	4673(7)	1930(8)	2032(8)	570(6)
6위	사인	당뇨병	당뇨병	자살	자살	자살
	빈도	2509(5)	3055(5)	995(4)	1560(6)	511(5)
7위	사인	운수사고	고혈압성질환	당뇨병	당뇨병	당뇨병
	빈도	2506(5)	2766(4)	963(4)	880(3)	353(3)
8위	사인	간질환	만성하기도질환	호흡기결핵	추락	익사
	빈도	2327(4)	2241(4)	563(2)	489(2)	180(2)
9위	사인	호흡기결핵	호흡기결핵	고혈압성질환	호흡기결핵	추락
	빈도	1060(2)	1495(2)	553(2)	461(2)	165(2)
10위	사인	폐렴	자살	만성하기도질환	익사	고혈압성질환
	빈도	858(2)	1284(2)	497(2)	459(2)	163(2)
M^2		583 (p<.0001)				

		2004년				
		무학	초등학교 졸업	중학교 졸업	고등학교 졸업	대학 이상
1위	사인	악성신생물	악성신생물	악성신생물	악성신생물	악성신생물
	빈도	13471(25)	22182(36)	9440(37)	12834(38)	6403(42)
2위	사인	뇌혈관질환	뇌혈관질환	뇌혈관질환	뇌혈관질환	뇌혈관질환
	빈도	12903(24)	11432(19)	3681(14)	4037(12)	1804(12)
3위	사인	심장질환	심장질환	심장질환	자살	자살
	빈도	5888(11)	5419(9)	2103(8)	3534(10)	1494(10)
4위	사인	만성하기도질환	당뇨병	자살	심장질환	심장질환
	빈도	3658(7)	4261(7)	2030(8)	2856(8)	1492(10)
5위	사인	당뇨병	간질환	간질환	운수사고	운수사고
	빈도	3516(7)	3197(5)	1926(8)	2396(7)	932(6)
6위	사인	고혈압성질환	만성하기도질환	당뇨병	간질환	당뇨병
	빈도	2493(5)	2986(5)	1541(6)	2208(7)	674(4)
7위	사인	폐렴	자살	운수사고	당뇨병	간질환
	빈도	1566(3)	2946(5)	1260(5)	1701(5)	618(4)
8위	사인	자살	운수사고	만성하기도질환	추락	만성하기도질환
	빈도	1456(3)	2367(4)	719(3)	665(2)	305(2)
9위	사인	운수사고	고혈압성질환	추락	만성하기도질환	추락
	빈도	1323(3)	1518(3)	427(2)	650(2)	247(2)
10위	사인	간질환	폐렴	호흡기결핵	호흡기결핵	고혈압성질환
	빈도	1240(2)	1085(2)	419(2)	447(1)	179(1)
M^2		1141 (p<.0001)				

※ ( ) : 각 교육 수준 범주에서 해당사인이 차지하고 있는 비율(%)

표2.8은 결혼 상태의 각 범주의 주요사망원인에 대한 빈도표이다. 결혼 상태에 따라 사망원인별로 유의한 차이가 있는지 알아보기 위하여 독립성 검정을 해본 결과 피어슨카이제곱통계량이 1995년과 2004년에서 모두 유의하였다. 즉 결혼 상태에 따라 사인별로 유의한 차이가 존재함을 의미한다.

사망자의 결혼 상태가 미혼이나 이혼일 경우의 자살로 인한 사망비율이 배우자가 있거나 사별인 경우보다 높았고 1995년에 비하여 2004년에 비율이 증가했다. 미혼은 사고사로 인한 사망이 많은 반면 나머지 결혼 상태의 범주에서는 질환사로 인한 사망이 많다. 이는 사망자의 연령과 연관을 지어

생각해 보아야 한다.

표2.8 결혼 상태별 사인순위

		1995년				2004년			
		미혼	유배우	이혼	사별	미혼	유배우	이혼	사별
1위	사인	운수사고	악성신생물	악성신생물	뇌혈관질환	악성신생물	악성신생물	악성신생물	악성신생물
	빈도	6267(31)	33653(34)	1008(23)	16539(30)	2700(18)	42177(41)	2572(28)	16352(27)
2위	사인	악성신생물	뇌혈관질환	간질환	악성신생물	자살	뇌혈관질환	간질환	뇌혈관질환
	빈도	2390(12)	17419(18)	635(14)	12289(22)	2502(16)	16223(16)	1187(13)	15292(25)
3위	사인	자살	간질환	뇌혈관질환	심장질환	운수사고	심장질환	자살	심장질환
	빈도	1706(8)	9130(9)	624(14)	6601(12)	2056(13)	8532(8)	1182(13)	7463(12)
4위	사인	간질환	운수사고	운수사고	고혈압성질환	출생전후기질환	자살	뇌혈관질환	당뇨병
	빈도	1224(6)	8656(9)	492(11)	4106(7)	1197(8)	5920(6)	1076(12)	4729(8)
5위	사인	심장질환	심장질환	심장질환	만성하기도질환	뇌혈관질환	당뇨병	심장질환	만성하기도질환
	빈도	1221(6)	8254(8)	327(7)	3326(6)	987(6)	5900(6)	770(8)	3768(6)
6위	사인	익사	당뇨병	자살	당뇨병	간질환	간질환	당뇨병	고혈압성질환
	빈도	1055(5)	4145(4)	241(5)	3042(6)	978(6)	5636(5)	546(6)	2985(5)
7위	사인	뇌혈관질환	고혈압성질환	당뇨병	간질환	심장질환	운수사고	운수사고	자살
	빈도	960(5)	3798(4)	200(4)	2145(4)	903(6)	4500(4)	519(6)	1793(3)
8위	사인	선천기형	만성하기도질환	호흡기결핵	운수사고	선천기형	만성하기도질환	추락	폐렴
	빈도	943(5)	3031(3)	175(4)	1899(3)	652(2)	4094(4)	191(2)	1783(3)
9위	사인	추락	자살	행동장애	호흡기결핵	익사	고혈압성질환	행동장애	간질환
	빈도	581(3)	2273(2)	131(3)	871(2)	462(3)	1678(2)	183(2)	1358(2)
10위	사인	호흡기결핵	호흡기결핵	고혈압성질환	폐렴	당뇨병	추락	만성하기도질환	추락
	빈도	513(3)	2126(2)	111(2)	740(1)	462(3)	1478(1)	181(2)	1218(2)
χ^2		49769 (p<.0001)				49323 (p<.0001)			

※ ( ) : 각 결혼 상태별 범주에서 해당사인이 차지하고 있는 비율(%)

2.3 수량화방법Ⅱ를 이용한 탐색적 자료분석

수량화방법은 林知己夫(Hayashi chikio)(1970)등에 의해 체계화된 이론으로서, 질적 데이터에 수량을 부여하여 중회귀분석, 주성분분석, 판별분석과 같은 다차원적 해석을 수행하는 다변량 해석 기법중 하나이다. 특히 수량화방법Ⅱ는 질적 변수인 외적기준과 독립변수의 각 범주에 수량화 값을 부여하는 것으로, 외적변수와 독립변수를 모두 가변수를 이용하여 표현한 뒤 외적기준 가변수들의 선형결합과 설명변수 가변수들의 선형결합 간의 상관계수를 최대화함으로써 모든 질적 범주에 수량화 값을 부여하는 기법이다.

앞 절에서 살펴본 빈도·교차분석결과 성, 연령, 결혼상태, 교육정도에 따라 사망원인과의 연관성이 존재한다는 알 수 있었는데 이를 구체적으로 확인하기 위하여 수량화방법Ⅱ를 이용한 정준상관분석을 실시하여 보도록 하자.

1995년부터 2004년까지 사망자를 대상으로 빈도분석결과 주요10대 사망원인을 찾을 수 있었다. 10대 사인과 관련 있는 위험인자를 식별하기 위하여 5개의 설명변량(주소지, 성별, 연령, 교육정도, 결혼상태)을 사용하였으며 이에 대한 범주표는 표2.9와 같다.

먼저 사망원인 자료의 수량화를 위하여 정준상관분석을 실시하여 보자.

SAS의 PROC CANCORR를 사용하여 얻은 원수량화값은 가변수 설정시 제외된 마지막 범주의 수량화 값이 모두 0이다. 하야시의 수량화 방법에서는 개별 범주의 주변빈도들을 가중치로한 가중평균을 사용하여 중심화 하여 재표현 한다. 또한 각 설명변량의 중요도(기여도)를 비교하는 지표로서 수량화 값의 범위 또는 수량화된 변수들 사이의 편상관을 쓰는데 본 논문에서는 설명변량 내 범주들의 수량화 값들 중에서 최대값과 최소값의 차이인 수량화 값의 범위를 사용 하도록 하겠다.

표2.9 수량화 분석을 이용한 범주표

	범주1	범주2	범주3	범주4	범주5	범주6	범주7	범주8	범주9	범주10
사망원인	악성신생물	뇌혈관질환	심장질환	운수사고	간질환	당뇨병	하기도질환	자살	고혈압	호흡기결핵
주소지	서울/부산/대구/인천/광주/대전/경기도/강원도/충청북도/충청남도/전라북도/전라남도/경상북도/경상남도/제주도 (14개 범주)									
성별	남	여								
연령	10대이하	20대	30대	40대	50대	60대	70대	80대이상		
교육수준	무학	초졸	중졸	고졸	대학이상					
결혼상태	미혼	유배우	이혼	사별						

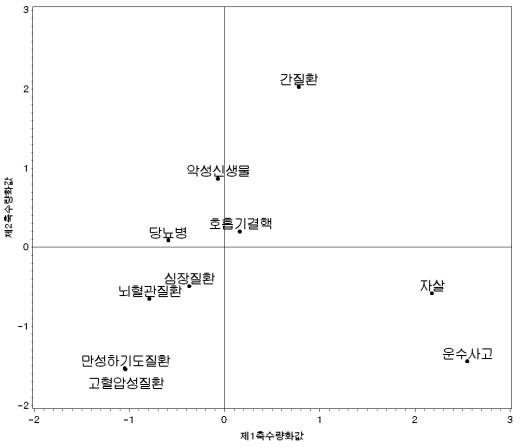
표2.10은 변수에 대한 빈도 및 수량화값, 범위를 나타낸 표이며 그림2.3은 표2.10을 범주별로 2차원 공간좌표에 플롯한 그림이다. 표2.10와 그림2.3으로부터 알 수 있듯이 외적기준의 제1축 수량화는 자살 혹은 운수사고와 같은 사고사와 질환에 의한 사망원인의 축으로 되어 있다. 이 중 자살과 운수사고는 양의 값을 가지고 나머지 사인은 음의 값을 가진다. 이에 가장 큰 관련을 갖는 설명변량은 연령으로 30대 이하 연령대는 사고사의 방향과 관련되어 있다. 그 다음으로 큰 영향을 갖는 설명변량은 성별과 주소지이며 여성일수록 또는 주소지가 대도시(서울, 부산, 대구, 인천, 광주)일수록 질환에 의한 사인에 관련이 있다. 그러나 연령을 제외하고는 그 영향력이 작으며 제1축의 정준상관계수는 0.51로 나타났다.

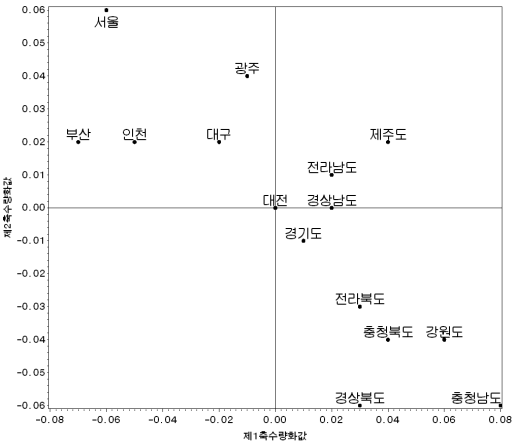
제2축 수량화는 만성하기도질환 및 고혈압성질환과 간질환의 대비로 되어 있다. 큰 값을 가질수록 만성질환의 성격을 가지고 있지만 정준상관계수가 0.29로 낮게 나타났다.

표2.10 변수들에 대한 수량화값과 범위

		빈도	제1축			제2축		
			원정준계수	수량화값	범위	원정준계수	수량화값	범위
사망 원인	악성신생물	557,372	-0.23	-0.07	3.6	0.67	0.87	3.57
	뇌혈관질환	345,870	-0.95	-0.79		-0.85	-0.65	
	심장질환	170,431	-0.53	-0.37		-0.69	-0.49	
	운수사고	122,249	2.40	2.55		-1.64	-1.44	
	간질환	110,124	0.62	0.78		1.83	2.03	
	당뇨병	101,280	-0.75	-0.59		-0.11	0.09	
	만성하기도질환	76,228	-1.20	-1.05		-1.73	-1.53	
	자살	76,005	2.02	2.18		-0.78	-0.58	
	고혈압성질환	50,249	-1.20	-1.04		-1.74	-1.54	
	호흡기결핵	31,707	0.00	0.16		0.00	0.20	
주소지	서울	255,985	-0.20	-0.06	0.15	0.15	0.06	0.12
	부산	133,723	-0.21	-0.07		0.01	0.02	
	대구	75,841	-0.12	-0.02		0.01	0.02	
	인천	71,471	-0.17	-0.05		0.00	0.02	
	광주	36,131	-0.09	-0.01		0.09	0.04	
	대전	36,302	-0.07	0.00		-0.04	0.00	
	경기도	256,966	-0.07	0.01		-0.08	-0.01	
	강원도	70,301	0.04	0.06		-0.19	-0.04	
	충청북도	66,549	0.01	0.04		-0.21	-0.04	
	충청남도	95,712	0.08	0.08		-0.28	-0.06	
	전라북도	95,074	-0.02	0.03		-0.15	-0.03	
	전라남도	119,312	-0.03	0.02		-0.02	0.01	
	경상북도	150,795	-0.02	0.03		-0.25	-0.06	
	경상남도	159,960	-0.03	0.02		-0.05	0.00	
	제주도	17,393	0.00	0.04		0.00	0.02	
성별	남자	981,367	0.32	0.07	0.17	0.30	0.03	0.08
	여자	660,148	0.00	-0.10		0.00	-0.05	
연령	10대이하	28,176	3.69	1.48	1.96	-1.42	-0.70	0.97

		빈도	제1축			제2축		
			원점준계수	수량화값	범위	원점준계수	수량화값	범위
	20대	43,053	3.84	1.55		-1.38	-0.68	
	30대	81,736	2.63	0.93		0.53	-0.13	
	40대	156,840	1.66	0.44		1.81	0.24	
	50대	241,804	1.06	0.13		1.92	0.27	
	60대	383,024	0.59	-0.11		1.49	0.15	
	70대	431,007	0.24	-0.29		0.82	-0.05	
	80대이상	275,875	0.00	-0.41		0.00	-0.28	
교육수준	무학	478,345	0.00	-0.03		-0.23	-0.05	
	초등학교졸업	558,834	0.10	0.02		0.02	0.02	
	중학교졸업	221,555	0.11	0.02	0.05	0.06	0.03	0.08
	고등학교졸업	269,271	0.11	0.02		-0.07	0.00	
	대학이상	113,510	0.00	-0.03		0.00	0.02	
결혼상태	미혼	125,374	0.19	0.10		0.09	-0.03	
	배우자있음	933,344	-0.04	-0.02	0.10	0.33	0.04	0.10
	이혼	57,893	0.15	0.08		0.25	0.01	
	사별	524,904	0.00	0.00		0.00	-0.06	
정준상관계수			0.51			0.29		





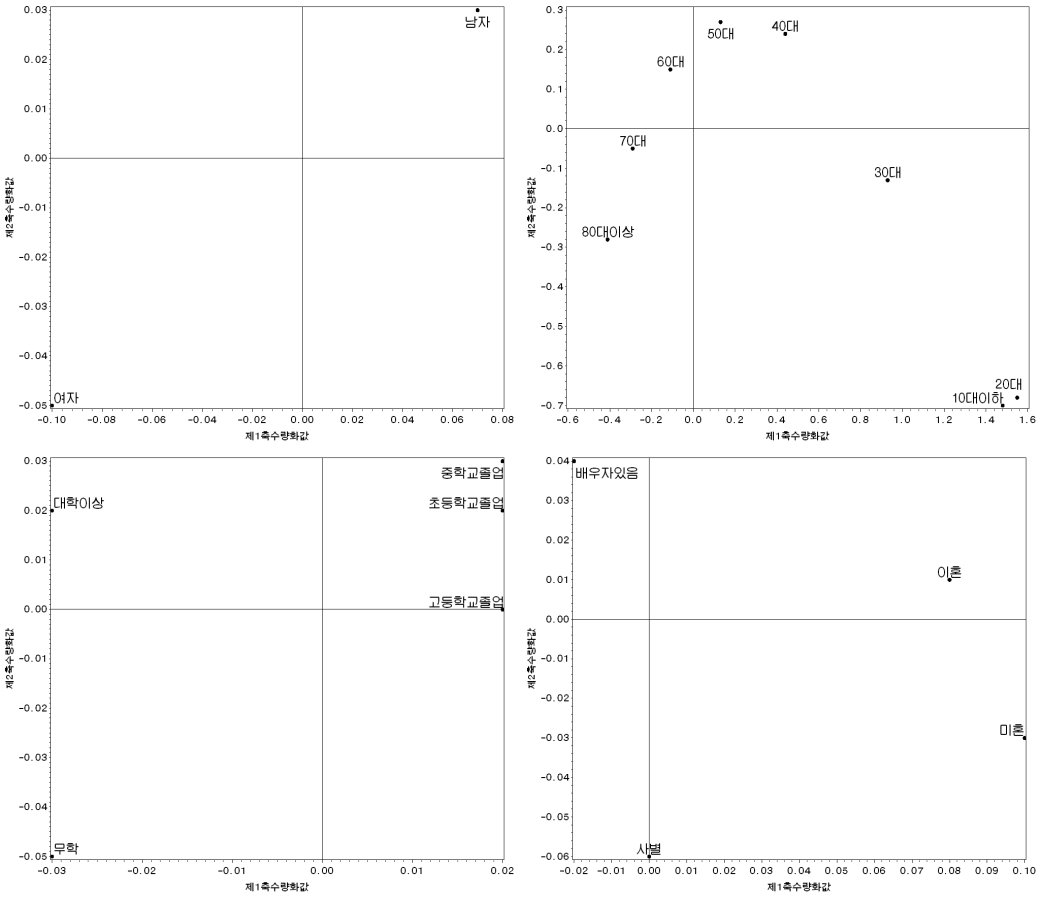


그림2.3 변수들에 대한 수량화 플롯

제3장 시계열 자료 분석

시계열분석은 시간의 흐름에 따라 변하는 현상을 관측함으로써 얻어지는 자료를 분석하여 설명할 수 있는 모형을 설정하고 미래를 예측하는 분석을 말한다(정동빈, 원태연; 2005). 즉 과거 자료의 관측값 만으로 자료의 형태를 설명하고, 자료에 영향을 미치는 경제·사회변수들 간의 관계를 고려하지 않고 단지 과거 움직임에 기초하여 규칙을 찾아내는 방법이라 말할 수 있다.

시계열 자료들을 분석하는 목적은 첫째로, 주어진 시계열 자료들을 생성하는 구조를 이해하고자 하는 것이고, 둘째로는 과거의 값들로부터 미래의 값을 예측(forecasting)하고자 하는데 있다. 이러한 목적을 달성하기 위해서는 우선 주어진 시계열 자료를 적합 시킬 수 있는 수학적 모형을 선택하여 모형의 모수를 추정한 후 자료의 적합성을 검토함으로써 선택된 모형을 시계열의 생성체계를 이해하는데 사용한다(이종협, 최기헌; 1994).

시계열 분석법에는 시계열회귀분석(time series regression analysis), 시계열분해기법(time series decomposition method), 지수평활법(exponetial smoothing method), Box-Jenkins의 ARIMA(autoregressive integrated moving average)분석 등이 있다.

제3장에서는 56항목의 각 사인별 월별 사망자수에 적합한 ARIMA모형을 제시하고 향후 사망 추이를 예측한다. 또한 우리나라 사망원인 중 1위에 해당하는 악성신생물을 IARC(International Agency for Research on Cancer)에서 발표하는 26개 악성신생물의 종류를 기준으로 우리나라에서는 드문 카포시 육종(Kaposi sarcoma)은 제외하고 흔한 담낭을 추가하여 26개 악성신생물에 의한 사망원인 자료에 대해 각각 ARIMA모형을 적합 시키고 이를 토대로

향후 2년간 사망자수를 파악한다.

마지막으로 26개 악성신생물중 최근에 급격한 증가 추이를 보이고 있는 대장암자료에 대하여 ARIMA모형, 분해모형 및 지수평활법으로 모형적합을 시도하여 최적모형을 선택한 후 향후 사망추이를 심층적으로 고찰한다.

3.1 분석모형

3.1.1 Box-Jenkins의 승법계절 ARIMA모형

Box-Jenkins의 ARIMA모형은 모형의 식별, 추정, 검진의 3단계에 걸쳐 진행되는 시계열 분석 및 예측기법이다. 이모형은 어떤 시계열에도 적용이 가능하나, 시계열의 구성요소가 시간의 흐름에 따라 빠르게 변동할 때 유용하다. 또한 Box-Jenkins의 시계열 분석방법은 확률모형에 근거를 두고 있다.

실제 계절적 특성을 가지고 있는 경제, 경영분야의 많은 자료들은 순수하게 계절적 요인만을 가지고 있는 경우는 흔치않다. 따라서 여러 해에 걸쳐 수집된 월별 또는 분기별 자료처럼 계절적 특성을 가지고 있는 시계열 자료에 대해서는 승법계절 시계열모형($ARIMA(p,d,q)\times(P,D,Q)_{12}$모형)을 사용하게 된다(이종협, 최기헌; 1994). 자료를 모형화 할 때 주어진 시계열이 비정상성을 가지고 있다면 계절차분과 일반차분을 통하여 정상화를 시킴으로써 모형을 구축할 수 있다.

시계열 y_t에서 일반 시계열 모형의 차수가 (p,d,q)이고, 계절주기가 s이며 계절시계열 모형의 차수가 (P,D,Q)라 하면 승법계절 ARIMA모형은 (1)과 같다.

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)a_t, \quad (1)$$

여기서 $\Phi_P(B^s)=1-\Phi_1 B^s-\Phi_2 B^{2s}- \cdots -\Phi_P B^{Ps}$와 $\Theta_Q(B^s)=1-\Theta_1 B^s-\Theta_2 B^{2s}- \cdots -\Theta_Q B^{Qs}$은 각각 계절 자기회귀와 이동평균 다항식이고, $\phi_p(B)=1-\phi_1 B-\phi_2 B^2- \cdots -\phi_p B^p$와 $\theta_q(B)=1-\theta_1 B-\theta_2 B^2- \cdots -\theta_q B^q$은 각각 일반 자기회귀와 이동평균 다항식으로 공통근을 가지고 있지 않다. a_t는 백색잡음과정(White Noise Process)을 따르는 확률변수로 $j>0$에 대해 $E(y_{t-j}\cdot a_t)=0$이라 가정한다.

3.1.2 지수평활법 (Exponential smoothing method)

지수평활법은 시계열의 구성요소가 시간에 따른 변동이 느리거나 규칙적인 형태를 보일 경우 사용하는 시계열 예측방법으로서 최근의 자료에 더 큰 가중치를 주고 과거로 갈수록 가중치를 지수적(exponentially)으로 줄여나가는 방식이다.

시계열이 계절성분을 가지고 움직인다면 Winters의 계절모형을 사용하는 것이 바람직하다. Winters의 계절모형은 계절요인의 변동 양상에 따라 두가지로 구분한다. 시계열이 계절성분을 가지고 움직이며 계절요인의 변동폭이 일정하다면 Winster의 가법계절모형(Winsters' additive seasonal model)을, 변동폭이 시간의 흐름에 따라 점차 커지는 경우 Winters의 승법계절모형(Winters' multiplicative seasonal model)을 사용한다. Winster의 계절모형은 시계열이 추세성분과 계절성분 및 불규칙 성분들의 합 또는 곱으로 구성되어 있다고 보고 각 성분들을 평활법에 의해 추정한 후 이를 이용하여 예측값을 구하는

방법이다.

$$\text{Winters의 가법계절모형} : y_t = T_t + S_t + I_t, \quad (2)$$

$$\text{Winters의 승법계절모형} : y_t = T_t \times S_t \times I_t, \quad (3)$$

여기서 T_t는 추세성분, S_t는 계절주기 s를 가지는 계절성분, I_t는 오차항으로서 불규칙성분에 해당된다. 선형추세를 가정하면 $T_t = \beta_0 + \beta_1 t$ 의 관계를 갖는다. 또한 각 성분에 대응되는 3개의 평활함수와 평활상수를 갖는데 평활상수는 0과 1사이의 값을 가지며 변화정도가 클수록 1에 가까운 값을 가진다.

지수평활법은 평활상수의 선택이 임의이고 특정모형 하에서만 최적이며, 예측구간을 구하기 어렵다는 단점이 있다. 그럼에도 불구하고 ARIMA모형에 비해 비교적 직관적이고 사용하기 편리하고 활용가치가 높다는 장점이 있다.

3.1.3 분해모형(Decomposition model)

분해기법은 시계열자료 y_t를 구성하고 있는 성분들이 결정적(deterministic)이며 서로 독립이라고 가정하고 y_t를 추세요인(T_t), 계절요인(S_t), 순환요인(C_t), 불규칙요인(I_t)으로 분해한 후 이를 이용하여 미래를 예측하는 방법이다.

시계열의 변동 요소들이 시계열에 가법적 또는 승법적으로 포함 되었느냐에 따라 분해모형을 두 가지 형태로 구분한다.

가법모형 : $y_t = T_t + S_t + C_t + I_t,$ (4)

승법모형 : $y_t = T_t \times S_t \times C_t \times I_t.$ (5)

주어진 시계열 자료에 모형 (4)과 (5)중 어느 것을 적용할 것인가는 시도표를 통해 추세와 계절성의 유무를 파악하여 결정한다. 일반적으로 시계열 자료의 변동이 규칙적이고 선형추세에 의존하지 않으며 계절 변동이 일정한 형태를 가지고 있는 경우는 가법모형이 적절하며, 시간이 경과함에 따라 일정한 추세를 가지고 선형적으로 증가하며 계절 변동이 커지거나 작아진다면 승법모형이 적절하다고 판단한다. 그러나 순환성분 C_t는 S_t보다 주기가 길며 이 주기를 찾는 것은 쉬운 일이 아니므로 분해모형 사용시 순환성분은 고려하지 않는 것이 일반적이다.

3.1.4 모형 선정 기준

주어진 시계열 자료에 여러 가지 후보모형이 존재할 경우 일반적으로 사용되는 모형 선택 기준은 잔차 $\hat{a_t}$에 근거한 Akaike(1973, 1974)의 AIC (Akaike's Information Criterion) 또는 Schwarz (1978)의 SBC(Schwartz's Baysian Criterion)통계량 등이 주로 사용되며 모형 선정시 AIC와 SBC값을 최소로 하는 모형을 선택한다.

시계열 y_t을 $ARIMA(p,q)$모형에 적합 시켰을 때 AIC와 SBC는 (6)과 (7)로 정의된다.

$$\text{AIC}=n\cdot\ln\widehat{\sigma^2}+2\cdot(p+q), \tag{6}$$

$$\text{SBC}=n\cdot\ln\widehat{\sigma^2}+2\cdot(p+q)\cdot\ln n, \tag{7}$$

여기서 n은 시계열로부터 계산될 수 있는 잔차 $\hat{a_t}$의 개수이며, $\hat{\sigma}^2$는 오차항의 분산 추정치이다.

반면에 예측이 자료 분석의 주목적인 경우 예측의 정도를 평가하는데 예측오차에 근거한 적합척도인 제곱근평균제곱오차(RMSE: Root Mean Square Error), 평균절대퍼센트오차(MAPE: Mean Absolute Percentage Error) 등을 사용하여 모형을 선택할 수 있다.

주로 사용되는 RMSE와 MAPE는 다음과 같이 정의한다.

$$RMSE=\sqrt{\frac{1}{n-k}\sum_{t=1}^{n}(y_t-\hat{y_t})^2}\quad , \tag{8}$$

$$MAPE=\frac{100}{n}\times\sum_{t=1}^{n}\left|\frac{(y_t-\hat{y_t})}{y_t}\right|\ (\%). \tag{9}$$

후보모형의 RMSE와 MAPE를 비교하여 값이 작은 모형을 최적모형으로 선정한다.

3.2 56항목의 ARIMA모형 적합

사망원인통계는 19개의 장, 103항목(WHO에서 유의하여 볼 필요성이 있는 사인으로 권고한 항목), 56항목(사인순위 선정을 위한 항목), 236항목(우리나

라에서 많이 발생하는 사인을 위주로 한 항목)별로 구분하여 제표하고 있다. 본 논문에서는 통계청에서 사망원인통계를 발표할 때 사용하는 56항목의 각 사인별로 ARIMA모형을 적합하고, 향후 사망추이를 예측하여 보도록 하겠다. 단, 10년간의 사망자수가 100명 이하인 사망원인 16개[1]는 분석에서 제외하였다. 특히 영아사망에 관한 사인(출생전후기질환, 선천기형)은 사망특성상 신고 되지 않는 경우가 많아 1999년부터 외부 행정기관 자료를 이용하여 보완함에 따라 99년 이전자료를 제거한 후 모형 적합 및 예측을 시행하였다.

각 사인별 최적 모형은 표3.1과 같다. 40개 사인의 최적모형을 적합한 결과 21개가 계절성을 띄고 있는 것으로 나타났으며, 기타 결핵은 계절에만 영향을 받는 것으로 나타났다.

특징적으로 ARIMA(0,1,1)모형이 8개 사인(출생전후기질환, 바이러스 간염, 죽상경화증, 사구체질환 및 세뇨관-사이질성 질환, 급성 류마티스열 및 만성 류마티스 심장질환, 영양실조, 기타 급성하기도 감염, 수막염)의 최적모형으로 가장 많은 사인에 적합 하는 모형으로 나타났다. 특히 영아관련 사인(선천기형, 영아급사증후군)은 AR(1)모형을 따르는 것으로 나타났고, 타살을 제외한 사고사(운수사고, 자살, 추락, 익사, 중독사고, 화재사고)는 계절적 영향을 받고 있었다. 운수사고와 추락은 봄과 가을에, 자살은 봄, 익사와 중독사고는 여름, 화재사고는 겨울에 사망자수가 가장 많았다.

부록A는 표3.1에서 제시된 모형을 토대로 사인별 향후 1년간의 사망자수를 예측한 결과를 나타낸 시도표이다. 40개 사인 중 9개 사인(악성신생물, 자살, 고혈압성 질환, 폐렴, 추락, 알쯔하이머병, 감염성 기원으로 추정되는 설사 및 위장염, 기타 창자 감염성 질환)에 의한 사망자수는 증가할 것으로

1) 콜레라(A00), 페스트(A20), 디프테리아(A36), 백일해(A37), 수막알균감염(A39), 주로 성행위로 전파되는 감염(A50-A64), 급성 회색질척수염(A80), 광견병(A82), 황열(A95), 홍역(B05), 말라리아(B50-B54), 리슈만편모충증(B55), 파동편모충증(B56-B57), 주혈흡충증(B65), 유산된임신(O00-O07), 간접산과적사망(O98-O99)

보이며, 6개사인(운수사고, 간질환, 당뇨병, 만성 하기도질환, 호흡기 결핵, 중독사고)에 의한 사망자수는 감소할 것으로 예측된다.

다음으로 우리나라 제1사망원인 악성신생물을 각 장기별로 나누어 살펴보자.

표3.1 사인별 최적 적합모형

순위	사인명	적합모형
1	악성신생물	ARIMA(0,1,1)×(1,0,0)$_{12}$ $(1-\underset{(0.07)}{0.64}B^{12})(1-B)y_t = a_t(1-\underset{(0.07)}{0.66}B)a_t$
2	뇌혈관질환	ARIMA(1,0,0)×(2,1,0)$_{12}$ $(1-\underset{(0.08)}{0.59}B)(1+\underset{(0.08)}{0.46}B^{12}+\underset{(0.09)}{0.58}B^{24})(1-B^{12})y_t = a_t$
3	심장질환	ARIMA(1,0,0)×(2,0,0)$_{12}$ $(1-\underset{(0.07)}{0.69}B)(1-\underset{(0.08)}{0.32}B^{12}-\underset{(0.09)}{0.33}B^{24})(y_t-\underset{(58.88)}{1442.77}) = a_t$
4	운수사고	ARIMA(1,0,1)×(2,1,0)$_{12}$ $(1-\underset{(0.06)}{0.91}B)(1+\underset{(0.10)}{0.60}B^{12}+\underset{(0.09)}{0.49}B^{24})(1-B^{12})\ln y_t = 0.02+(1-\underset{(0.12)}{0.53}B)a_t$
5	간질환	ARIMA(0,1,1)×(2,0,0)$_{12}$ $(1-\underset{(0.08)}{0.32}B^{12}-\underset{(0.09)}{0.69}B^{24})(1-B)y_t = (1-\underset{(0.05)}{0.82}B)\,a_t$
6	당뇨병	ARIMA(1,0,1)×(2,1,0)$_{12}$ $(1-\underset{(0.02)}{0.98}B)(1+\underset{(0.09)}{0.46}B^{12}+\underset{(0.10)}{0.45}B^{24})(1-B^{12})\ln y_t = (1-\underset{(0.08)}{0.69}B)\,a_t$
7	만성 하기도질환	ARIMA(0,1,0)×(1,0,0)$_{12}$ $(1-B)(1-\underset{(0.09)}{0.37}B^{12})\ln y_t = a_t$
8	자살	ARIMA(1,1,0)×(1,1,1)$_{12}$ $(1+\underset{(0.09)}{0.28}B)(1+\underset{(0.12)}{0.25}B^{12})(1-B)(1-B^{12})\ln y_t = (1-\underset{(0.14)}{0.74}B^{12})a_t$
9	고혈압성질환	ARIMA(1,0,0)×(0,0,1)$_{12}$ $(1-\underset{(0.03)}{0.92}B)(y_t-\underset{(66.64)}{466.33}) = (1+\underset{(0.09)}{0.28}B^{12})a_t$
10	호흡기결핵	ARIMA(1,1,1)×(1,0,0)$_{12}$

순위	사인명	적합모형
		$(1-0.34B)(1-0.36B^{12})(1-B)y_t=(1-0.89B)a_t$ (0.11) (0.09) (0.06)
11	폐렴	ARIMA(1,1,1)×(1,0,1)$_{12}$ $(1-0.40B)(1-0.98B^{12})(1-B)\ln y_t=(1-0.81B)(1-0.88B^{12})a_t$ (0.14) (0.07) (0.09) (0.24)
12	추락	ARIMA(0,1,1)×(1,0,1)$_{12}$ $(1-0.97B^{12})(1-B)y_t=(1-0.78B)(1-0.80B^{12})a_t$ (0.05) (0.06) (0.16)
13	익사	ARIMA(3,1,0)×(0,1,1)$_{12}$ $(1+0.54B+0.47B^2+0.36B^3)(1-B)(1-B^{12})\ln y_t=(1-0.80B^{12})a_t$ (0.09) (0.09) (0.09) (0.12)
14	정신활성물질 사용에 의한 정신 및 행동장애	ARIMA(0,1,1)×(0,0,1)$_{12}$ $(1-B)y_t=(1-0.82B)(1+0.30B^{12})a_t$ (0.05) (0.09)
15	패혈증	ARIMA(1,1,0) $(1+0.39B)(1-B)y_t=a_t$ (0.08)
16	출생전후기질환	ARIMA(0,1,1) $(1-B)y_t=(1-0.53B)a_t$ (0.10)
17	선천기형	AR(1) $(1-0.50B)(y_t-74.29)=a_t$ (0.11) (2.42)
18	타살	ARMA(1,1) $(1-0.84B)(y_t-69.79)=(1-0.62B)a_t$ (0.11) (2.31) (0.16)
19	중독사고	ARIMA(0,1,1)×(1,0,0)$_{12}$ $(1-0.38B^{12})(1-B)y_t=(1-0.40B)a_t$ (0.09) (0.09)
20	위 및 십이지장 궤양	ARIMA(0,1,1) $(1-B)y_t=(1-0.73B)a_t$ (0.06)
21	화재사고	ARIMA(1,0,0)×(0,1,1)$_{12}$ $(1-0.38B)(y_t-54.08)=(1+0.26B^{12})a_t$ (0.09) (4.27) (0.10)
22	바이러스 간염	ARIMA(0,1,1) $(1-B)\ln y_t=(1-0.72B)a_t$ (0.06)
23	알쯔하이머병	ARIMA(1,1,0)×(0,0,1)$_{12}$ $(1+0.44B)(1-B)y_t=(1+0.42B^{12})a_t$ (0.08) (0.10)

순위	사인명	적합모형
24	빈혈	ARMA(1,1) $(1-0.92B)(y_t-24.44)=(1-0.81B)a_t$ (0.09) (1.12) (0.13)
25	죽상경화증 (동맥경화증)	ARIMA(0,1,1) $(1-B)y_t=(1-0.56B)a_t$ (0.08)
26	나머지 특정 감염성 및 기생충성 질환	ARIMA(1,1,1) $(1-0.31B)(1-B)y_t=(1-0.89B)a_t$ (0.11) (0.06)
27	사구체질환 및 세뇨관-사이질성 질환	ARIMA(0,1,1) $(1-B)y_t=(1-0.82B)a_t$ (0.05)
28	급성 류마티스열 및 만성 류마티스 심장질환	ARIMA(0,1,1) $(1-B)y_t=(1-0.65B)a_t$ (0.07)
29	기타 결핵	ARIMA(0,0,1)$_{12}$ $(y_t-14.72)=(1+0.19B)a_t$ (0.44) (0.09)
30	영양실조	ARIMA(0,1,1) $(1-B)y_t=(1-0.55B)a_t$ (0.08)
31	기타 급성 하기도 감염	ARIMA(0,1,1) $(1-B)y_t=(1-0.37B)a_t$ (0.09)
32	수막염	ARIMA(0,1,1) $(1-B)y_t=(1-0.66B)a_t$ (0.07)
33	감염성기원으로 추정되는 설사 및 위장염	ARMA(1,1) $(1-0.93B)(y_t-10.36)=(1-0.67B)a_t$ (0.05) (1.82) (0.10)
34	영아급사증후군	AR(1) $(1-0.39B)(y_t-9.01)=a_t$ (0.08) (0.58)
35	인플루엔자	ARIMA(1,0,0)×(1,0,0)$_{12}$ $(1-0.49B)(1-0.28B^{12})(y_t-8.89)=a_t$ (0.08) (0.09) (2.03)
36	기타 창자 감염성질환	ARIMA(1,0,0)×(0,0,1)$_{12}$ $(1-0.44B)(y_t-5.02)=(1+0.18B^{12})a_t$ (0.08) (0.55) (0.09)
37	기타 직접산과적 사망	White Noise $y_t=4.87+a_t$ (0.19)
38	인체 면역 결핍 바이	ARIMA(1,1,1)

순위	사인명	적합모형
	러스병	$(1+0.21B)(1-B)y_t=(1-0.81B)a_t$ (0.10) (0.06)
39	기타 절지동물 매개의 바이러스열 및 바이러스 출혈열	ARIMA(1,0,0)×(0,1,1)$_{12}$ $(1-0.35B)(1-B^{12})y_t=(1-0.60B^{12})a_t$ (0.09) (0.10)
40	파상풍	White Noise $y_t=0.87+a_t$ (0.09)

※ ( ) : 표준오차

3.3 악성신생물의 ARIMA모형 적합

악성신생물은 우리나라 국민의 사망원인 1위로서 국민건강을 위협하는 가장 중대한 원인중 하나이다. 특히 노령화와 생활양식의 서구화, 환경적인 요인(흡연, 음주, 감염 등)으로 인하여 악성신생물의 발생 및 사망이 크게 증가하고 있어 이에 대한 양상을 살펴보아 추이를 예측하고자 한다.

그림3.1은 악성신생물의 시도표인데, 점차 증가하는 추세에 있으며 1월을 전후한 겨울에 사망이 적은 것으로 보인다. 악성신생물로 인한 사망률은 1983년 인구10만명당 70.5명이었으나 1989년에 105명으로 급격히 증가하였으며 1990년 110.4명에서 2006년 현재 134.5명으로 증가속도는 둔화되었으나 지속적으로 증가하는 추세에 있다.

악성신생물의 여러 가지 종류 중에서 IARC(International Agency for Research on Cancer)에서 발표하는 26개 암종을 기준으로 우리나라에서 드문 카포시육종(Kaposi sarcoma)을 제외하고, 흔한 담낭(Gallbladder etc.)를 추가하여 분석에 사용하였다.

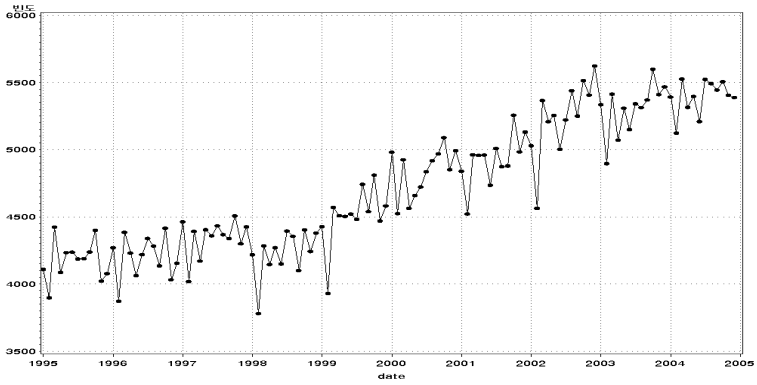


그림3.1 악성신생물의 시계열 시도표

각 장기별 악성신생물의 최적 모형은 표3.2와 같다. 7개 장기(위, 폐, 간, 대장, 담낭, 구강, 기타인두)에서 계절성을 띄고 있었으며 췌장, 백혈병, 비호지킨 림프종, 자궁경부, 후두, 방광, 전립선, 난소, 신장, 다발성 골수증, 피부 흑색종, 비인두, 자궁체부 13개 장기의 최적모형이 ARIMA(0,1,1)[2]으로 적합되었다.

방광암의 경우 ARIMA(0,1,1)모형 : $y_t = y_{t-1} + a_t - 0.75a_{t-1}$, (10)

이는 차분되어진 계열이 MA(1)을 따르고 있음을 나타내는 것이다. 즉, 현재시점(y_t)와 바로 전시점(y_{t-1})의 편차를 설명하기 위해 y_{t-1}만이 정보를 가지고 있다는 것이다. 다시 말하면, 방광암 시계열은 관측시점 전달의 사망자수(y_{t-1})와 관측시점과 관측시점 전달의 불규칙변동의 가중평균으로 ($a_t - 0.75a_{t-1}$)이루어져 있다.

2) 백혈병, 뇌 및 중추신경계, 후두, 전립선, 신장, 피부 흑색종은 ARMA(1,1)으로 판단가능하나 모형적합통계량인 RMSE와 MAPE기준에 의하여 ARIMA(0,1,1)로 모형적합 하였다.

부록B는 표3.2에서 제시된 모형으로 발생 부위별 향후 1년의 사망자수를 예측한 결과를 나타낸 시도표이다. 26가지 악성신생물 중에서 폐암으로 인한 사망자수가 가장 높은 증가율을 보이고 있었으며 그 다음으로 대장암, 췌장암, 전립선암에 의한 사망자수가 높은 증가율을 보인 반면, 후두암으로 인한 사망자수는 감소하는 추세를 가지고 있었다.

표3.2 악성신생물의 최적 적합모형

순위	발생부위	빈도(명)	적합모형
1	위	115,705	ARIMA(3,0,0)×(2,0,0)$_{12}$ $(1-\underset{(0.09)}{0.10}B-\underset{(0.09)}{0.24}B^2-\underset{(0.09)}{0.23}B^3)(1-\underset{(0.09)}{0.25}B^{12}-\underset{(0.10)}{0.35}B^{24})(y_t-\underset{(15.79)}{966.66})=a_t$
2	폐	109,230	ARIMA(0,1,1)×(0,1,1)$_{12}$ $(1-B)(1-B^{12})y_t=(1-\underset{(0.06)}{0.81}B)(1-\underset{(0.10)}{0.71}B^{12})a_t$
3	간	101,974	ARIMA(0,1,1)×(0,0,1)$_{12}$ $(1-B)\,y_t=(1-\underset{(0.05)}{0.81}B)(1+\underset{(0.09)}{0.22}B^{12})a_t$
4	대장	40,565	ARIMA(0,1,1)×(0,1,1)$_{12}$ $(1-B)(1-B^{12})y_t=(1-\underset{(0.06)}{0.79}B)(1-\underset{(0.10)}{0.77}B^{12})a_t$
5	췌장	25,487	ARIMA(0,1,1) $(1-B)\,y_t=(1-\underset{(0.06)}{0.77}B)\,a_t$
6	담낭	25,390	ARIMA(2,1,0)×(0,0,1)$_{12}$ $(1+\underset{(0.09)}{0.72}B+\underset{(0.09)}{0.25}B^{12})\,(1-B)y_t=(1+\underset{(0.09)}{0.32}B^{12})a_t$
7	식도	14,636	White Noise $y_t=\underset{(1.15)}{121.97}+a_t$
8	백혈병	13,845	ARIMA(0,1,1) $(1-B)\,y_t=(1-\underset{(0.04)}{0.91}B)a_t$

순위	발생부위	빈도(명)	적합모형
9	유방	11,750	ARIMA(1,1,1) $(1+\underset{(0.10)}{0.27B})\,(1-B)y_t = (1-\underset{(0.07)}{0.77B})a_t$
10	뇌 및 중추신경계	9,681	AR(2) $(1-\underset{(0.08)}{0.31B}-\underset{(0.08)}{0.43B^2})(\,y_t-\underset{(3.98)}{80.90}) = a_t$
11	비호지킨 림프종	8,327	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.07)}{0.61B})\,a_t$
12	자궁경부	7,912	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.06)}{0.74B})a_t$
13	후두	7,470	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.05)}{0.88B})\,a_t$
14	방광	7,098	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.06)}{0.75B})a_t$
15	전립선	5,346	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.07)}{0.70B})a_t$
16	난소	5,143	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.05)}{0.82B})a_t$
17	신장	4,875	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.05)}{0.85B})a_t$
18	구강	4,069	ARIMA(0,1,1)×(1,0,0)$_{12}$ $(1+\underset{(0.09)}{0.24B^{12}})\,(1-B)y_t = (1-\underset{(0.08)}{0.56B})a_t$
19	다발성 골수증	3,056	ARIMA(0,1,1) $(1-B)\,y_t = (1-\underset{(0.06)}{0.80B})\,a_t$
20	갑상선	2,415	ARIMA(4,1,0)

순위	발생부위	빈도(명)	적합모형
			$(1+0.72B+0.63B^2+0.51B^3+0.32B^4)(1-B)\,y_t = a_t$ (0.09) (0.10) (0.10) (0.09)
21	기타인두	2,053	ARIMA(1,1,1)×(0,0,1)$_{12}$ $(1+0.22B)(1-0.32B^{12})(1-B)y_t = (1-0.77B)a_t$ (0.11) (0.09) (0.07)
22	피부흑색종	947	ARIMA(0,1,1) $(1-B)\,y_t = (1-0.90B)\,a_t$ (0.04)
23	비인두	923	ARIMA(0,1,1) $(1-B)\,y_t = (1-0.84B)\,a_t$ (0.05)
24	자궁체부	660	ARIMA(0,1,1) $(1-B)\,y_t = (1-0.75B)a_t$ (0.06)
25	호지킨	230	ARIMA(6,1,0) $(1-0.91B+0.85B^2+0.59B^3+0.60B^4+0.27B^5+0.39B^6)(1-B)y_t = a_t$ (0.09) (0.12) (0.14) (0.14) (0.13) (0.10)
26	고환	150	White Noise $y_t = 1.25 + a_t$ (0.11)

※ ( ) : 표준오차

3.4 대장암에 대한 다양한 모형 적합

대장암은 결장과 직장에 생기는 악성 종양을 말하며, 암이 발생하는 위치에 따라 결장에 생기는 암을 결장암, 직장에 생기는 암을 직장암이라고 하고, 이를 통칭하여 대장암 혹은 결장직장암이라고 한다(국립암센터 홈페이지). 남녀에 상관없이 비슷한 사망률을 보이고 있으며, 연령이 높아질수록

사망률은 높아지고 비만과 밀접한 연관이 있다고 알려져 있다.

암으로 인한 전체 사망자 중에서 4위를 차지하고 있으며, 최근 식생활이 서구화됨에 따라 대장암 발생·사망률이 현저하게 증가하고 있다(그림3.2참고). 1980년대 말에는 인구 10만명당 3명이었으나 2002년에는 10명으로 급격히 증가하였다.

이제 ARIMA모형을 포함한 여러 다른 시계열분석법을 적용하여 어느 분석법이 예측에 가장 적합한지 살펴보자.

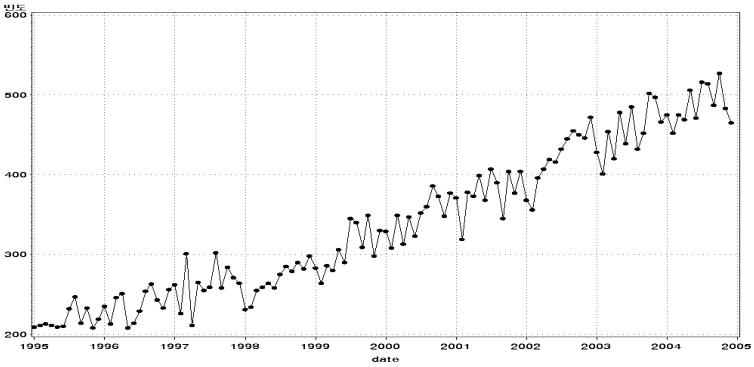


그림3-2 대장암으로 인한 사망자수 시도표

3.4.1 ARIMA모형

대장암의 월별 시계열 자료에 ARIMA모형을 적합하기 위하여 시도표를 그려본 결과(그림3-2) 시간이 경과함에 따라 사망자수가 증가하는 추세를 보이고 있다. 또한 표3.3과 같이 원계열의 자기상관이 시차가 커지더라도 사라지지 않고 서서히 줄어드는 형태이므로 차분(differencing)을 통하여 정상화

시켜 분석을 시도한다. 차분된 계열을 $ARIMA(p,d,q)\times(P,D,Q)_{12}$를 적합 시키기 위하여 SAS의 PROC ARIMA를 수행한 결과 표본자기상관함수(SACF)와 표본부분자기상관함수(SPACF)는 표3.4와 같다. 또한, 일반 차분된 계열의 표본자기상관함수(SACF)는 계절주기 12의 배수에 해당되는 시차를 따라 아주 느린 감소를 보여주고 있으므로 계절차분을 통하여 계절을 정상화시켜주는 것이 바람직하다. 이에 대한 SACF와 SPACF는 표3.5와 같다.

표3.3 원계열의 표본자기상관함수(SACF)

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	8653.673	1.00000	|**********|	0
1	8077.806	0.93345	|********** . |	0.091287
2	7996.837	0.92410	|********** . |	0.151181
3	7703.176	0.89016	|********** . |	0.192583
4	7452.725	0.86122	|********** . |	0.224265
5	7308.192	0.84452	|********** . |	0.250313
6	6966.092	0.80499	|********** . |	0.273026
7	6913.027	0.79885	|********** . |	0.292136
8	6660.737	0.76970	|********** . |	0.309806
9	6500.347	0.75117	|********** . |	0.325351
10	6364.650	0.73549	|********** . |	0.339496
11	6154.478	0.71120	|********** . |	0.352524
12	6086.909	0.70339	|********** . |	0.364285
13	5750.168	0.66448	|********** . |	0.375432
14	5578.796	0.64467	|********** . |	0.385108
15	5277.696	0.60988	|********** . |	0.393999
16	5070.581	0.58595	|********** . |	0.401789
17	4953.543	0.57242	|********** . |	0.408847
18	4525.526	0.52296	|********** . |	0.415472
19	4534.840	0.52404	|********** . |	0.420922
20	4225.318	0.48827	|********** . |	0.426324
21	4143.254	0.47879	|********** . |	0.430959
22	3981.911	0.46014	|********* . |	0.435369
23	3792.571	0.43826	|********* . |	0.439403
24	3758.307	0.43430	|********* . |	0.443031

표3.4 일반 차분된 계열의 SACF와 SPACF

Period(s) of Differencing	1
Mean of Working Series	2.151261
Standard Deviation	29.68848
Number of Observations	119
Observation(s) eliminated by differencing	1

Autocorrelations

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	881.406	1.00000	|||||||||||||||||||||	0
1	-521.048	-.59116	|***********|||||||||||	0.091670
2	163.952	0.18601	|||||||||||||||****||||||||||	0.119485
3	-8.173375	-.00927	|||||||||||||||||||||||||||||	0.121894
4	-143.451	-.16275	||||||||||||||***||||||||||||	0.121900
5	192.450	0.21834	|||||||||||||||****||||||||||	0.123713
6	-224.151	-.25431	|||||||||||||*****|||||||||||	0.126910
7	174.869	0.19840	|||||||||||||||****||||||||||	0.131122
8	-93.904158	-.10654	||||||||||||||**|||||||||||||	0.133621
9	-15.422322	-.01750	|||||||||||||||||||||||||||||	0.134333
10	67.729523	0.07684	|||||||||||||||**||||||||||||	0.134352
11	-161.467	-.18319	|||||||||||||****||||||||||||	0.134721
12	293.744	0.33327	||||||||||||||||******|||||||	0.136798
13	-229.329	-.26019	||||||||||||*****||||||||||||	0.143459
14	156.097	0.17710	|||||||||||||||****||||||||||	0.147371
15	-40.225397	-.04564	||||||||||||||||*||||||||||||	0.149149
16	-120.798	-.13705	|||||||||||||***|||||||||||||	0.149266
17	257.994	0.29271	|||||||||||||||*****|||||||||	0.150320
18	-380.160	-.43131	||||||||*******||||||||||||||	0.155036
19	302.228	0.34289	|||||||||||||||******||||||||	0.164811
20	-161.347	-.18306	||||||||||||****|||||||||||||	0.170700
21	17.352081	0.01969	|||||||||||||||||||||||||||||	0.172342
22	70.040117	0.07946	|||||||||||||||**||||||||||||	0.172361
23	-164.248	-.18635	||||||||||||****|||||||||||||	0.172669
24	342.531	0.38862	|||||||||||||||*******|||||||	0.174350
25	-283.983	-.32219	|||||||||||******||||||||||||	0.181484
26	157.203	0.17835	|||||||||||||||****||||||||||	0.186228
27	-67.864382	-.07700	||||||||||||||**|||||||||||||	0.187658
28	-15.956376	-.01810	|||||||||||||||||||||||||||||	0.187924
29	101.782	0.11548	|||||||||||||||**||||||||||||	0.187938
30	-230.963	-.26204	|||||||||||*****|||||||||||||	0.188533
31	234.831	0.26643	|||||||||||||||*****|||||||||	0.191570
32	-192.854	-.21880	||||||||||||****|||||||||||||	0.194658
33	45.322163	0.05142	||||||||||||||||*||||||||||||	0.196714
34	117.200	0.13297	||||||||||||||||***||||||||||	0.196827
35	-222.905	-.25290	|||||||||||*****|||||||||||||	0.197581
36	286.086	0.32458	|||||||||||||||******||||||||	0.200282
37	-219.528	-.24907	|||||||||||*****|||||||||||||	0.204655
38	169.845	0.19270	|||||||||||||||****||||||||||	0.207186
39	-109.492	-.12422	|||||||||||||**||||||||||||||	0.208687
40	48.880877	0.05546	|||||||||||||||*|||||||||||||	0.209308
41	42.468213	0.04818	|||||||||||||||*|||||||||||||	0.209431
42	-175.214	-.19879	|||||||||||||****||||||||||||	0.209524
43	176.763	0.20055	|||||||||||||||****||||||||||	0.211103
44	-147.276	-.16709	|||||||||||||***|||||||||||||	0.212698
45	65.507837	0.07432	|||||||||||||||*|||||||||||||	0.213798
46	62.802680	0.07125	|||||||||||||||*|||||||||||||	0.214015
47	-175.996	-.19968	|||||||||||||****||||||||||||	0.214215
48	277.257	0.31456	|||||||||||||||*****|||||||||	0.215773

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.59116	|***********|||||||||||
2	-0.25126	|||||||||||*****|||||||||||
3	-0.03317	|||||||||||||||*|||||||||||
4	-0.22667	|||||||||||*****|||||||||||
5	-0.00701	|||||||||||||||||||||||||||
6	-0.18560	|||||||||||||||****||||||||
7	-0.06365	||||||||||||||||*||||||||||
8	-0.07656	||||||||||||||||**|||||||||
9	-0.13011	||||||||||||||***||||||||||
10	-0.10165	|||||||||||||||**||||||||||
11	-0.25427	|||||||||||*****|||||||||||
12	0.10674	|||||||||||||||**||||||||||
13	0.02257	|||||||||||||||||||||||||||
14	0.09047	|||||||||||||||**||||||||||
15	0.09527	|||||||||||||||**||||||||||
16	-0.06089	||||||||||||||||*||||||||||
17	0.20365	|||||||||||||||****||||||||
18	-0.16228	||||||||||||||***||||||||||
19	-0.04605	||||||||||||||||*||||||||||
20	-0.05450	||||||||||||||||*||||||||||
21	-0.05371	||||||||||||||||*||||||||||
22	-0.07001	||||||||||||||||*||||||||||
23	-0.14190	||||||||||||||***||||||||||
24	0.22029	||||||||||||||||****|||||||
25	0.10518	||||||||||||||||**|||||||||
26	0.07011	||||||||||||||||*||||||||||
27	-0.02277	|||||||||||||||||||||||||||
28	0.08382	||||||||||||||||**|||||||||
29	0.04642	||||||||||||||||*||||||||||
30	-0.05627	||||||||||||||||*||||||||||
31	-0.00492	|||||||||||||||||||||||||||
32	-0.07444	||||||||||||||||*||||||||||
33	-0.10071	||||||||||||||||**|||||||||
34	0.05190	||||||||||||||||*||||||||||
35	-0.06756	||||||||||||||||*||||||||||
36	-0.12044	||||||||||||||||**|||||||||
37	-0.06069	||||||||||||||||*||||||||||
38	0.03912	|||||||||||||||||*|||||||||
39	-0.09746	||||||||||||||||**|||||||||
40	0.06617	||||||||||||||||*||||||||||
41	-0.00081	|||||||||||||||||||||||||||
42	0.06884	|||||||||||||||||*|||||||||
43	-0.02789	||||||||||||||||*||||||||||
44	0.00960	|||||||||||||||||||||||||||
45	-0.02952	||||||||||||||||*||||||||||
46	0.10766	||||||||||||||||**|||||||||
47	-0.06890	||||||||||||||||*||||||||||
48	-0.02921	||||||||||||||||*||||||||||

표3.5 일반 및 계절 차분된 계열의 SACF와 SPACF

Period(s) of Differencing	1,12
Mean of Working Series	-0.25234
Standard Deviation	34.35496
Number of Observations	107
Observation(s) eliminated by differencing	13

Autocorrelations

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	1180.263	1.00000	||********************|	0
1	-651.487	-.55198	|**********| . |	0.096674
2	86.129476	0.07297	| . |* . |	0.122641
3	34.949040	0.02961	| . |* . |	0.123046
4	-44.404774	-.03762	| . *| . |	0.123113
5	-55.462666	-.04699	| . *| . |	0.123220
6	147.948	0.12535	| . |*** . |	0.123388
7	-133.114	-.11278	| . **| . |	0.124572
8	85.724211	0.07263	| . |* . |	0.125523
9	-8.960723	-.00759	| . | . |	0.125915
10	-130.001	-.11015	| . **| . |	0.125919
11	410.802	0.34806	| . |******* |	0.126817
12	-677.515	-.57404	|**********| . |	0.135451
13	353.643	0.29963	| . |****** |	0.156544
14	18.881790	0.01600	| . | . |	0.161815
15	9.676461	0.00820	| . | . |	0.161830
16	-96.735349	-.08196	| . **| . |	0.161834
17	158.056	0.13392	| . |*** . |	0.162221
18	-198.429	-.16812	| . ***| . |	0.163251
19	113.725	0.09636	| . |** . |	0.164861
20	3.432417	0.00291	| . | . |	0.165387
21	22.681800	0.01922	| . | . |	0.165387
22	-75.024621	-.06357	| . *| . |	0.165408
23	29.576234	0.02506	| . |* . |	0.165636
24	178.116	0.15091	| . |*** . |	0.165672
25	-170.535	-.14449	| . ***| . |	0.166952
26	-16.789164	-.01422	| . | . |	0.168116
27	14.522670	0.01230	| . | . |	0.168128
28	20.676149	0.01752	| . | . |	0.168136
29	-39.956029	-.03385	| . *| . |	0.168153
30	40.609015	0.03441	| . |* . |	0.168217
31	29.405258	0.02491	| . | . |	0.168282
32	-69.922964	-.05924	| . *| . |	0.168317
33	-36.682927	-.03108	| . *| . |	0.168512
34	138.050	0.11697	| . |** . |	0.168565
35	-77.108084	-.06533	| . *| . |	0.169322
36	-126.090	-.10683	| . **| . |	0.169557
37	142.462	0.12070	| . |** . |	0.170185
38	15.781125	0.01337	| . | . |	0.170984
39	-58.851665	-.04986	| . *| . |	0.170993
40	93.508645	0.07923	| . |** . |	0.171129
41	-109.428	-.09271	| . **| . |	0.171472
42	82.308573	0.06974	| . |* . |	0.171940
43	-82.883131	-.07022	| . *| . |	0.172204
44	66.904378	0.05669	| . |* . |	0.172471
45	2.108080	0.00179	| . | . |	0.172645
46	-11.469072	-.00972	| . | . |	0.172645
47	-49.733358	-.04214	| . *| . |	0.172650
48	172.880	0.14648	| . |*** . |	0.172746

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.55198	|**********| . |
2	-0.33325	|*******| . |
3	-0.16281	| .***| . |
4	-0.11805	| . **| . |
5	-0.18357	| ****| . |
6	-0.01057	| . | . |
7	-0.05497	| . *| . |
8	0.00314	| . | . |
9	0.03531	| . |* . |
10	-0.13063	| .***| . |
11	0.36455	| . |******* |
12	-0.37005	|*******| . |
13	-0.30888	|******| . |
14	-0.16193	| .***| . |
15	0.04429	| . |* . |
16	-0.04493	| . *| . |
17	-0.06958	| . *| . |
18	0.00867	| . | . |
19	-0.04567	| . *| . |
20	0.04178	| . |* . |
21	0.14985	| . |***. |
22	-0.16261	| .***| . |
23	0.20148	| . |**** |
24	0.07749	| . |** . |
25	-0.01345	| . | . |
26	-0.09030	| . **| . |
27	0.05007	| . |* . |
28	-0.04370	| . *| . |
29	-0.07906	| . **| . |
30	-0.13772	| .***| . |
31	0.02450	| . | . |
32	0.00755	| . | . |
33	0.06545	| . |* . |
34	-0.08367	| . **| . |
35	0.16911	| . |***. |
36	0.00750	| . | . |
37	-0.08340	| . **| . |
38	-0.13456	| .***| . |
39	0.00966	| . | . |
40	0.07648	| . |** . |
41	-0.07998	| . **| . |
42	-0.04657	| . *| . |
43	0.02832	| . |* . |
44	-0.01061	| . | . |
45	0.01508	| . | . |
46	-0.01808	| . | . |
47	0.09199	| . |** . |
48	0.12879	| . |***. |

일반적으로 차분되는 시계열의 표본 평균과 그것의 표준 편차를 비교하여 절편항의 포함여부를 결정한다. 표본 평균을 표준 편차로 나눈 값의 절대치

가 2보다 클 경우 절편항을 모형에 포함시키는데 대장암으로 인한 사망자수 시계열 자료는 $|\sqrt{108}(29.9)/26.2|$=11.8로 2보다 크므로 절편항은 모형에 포함시켜야 하지만 최우추정법에 의한 추정의 결과 유의하지 않아 최종적으로 모형에서 제거하였다. SACF는 시차 1, 12에서 강한 자기상관관계를 보이는 반면에 SPACF는 시차 1, 12, …을 따라 지수적으로 감소하며 12의 배수에 해당되는 추세를 따라서는 감소하는 패턴을 보이므로 주어진 시계열은 ARIMA(0,1,1)×$(0,1,1)_{12}$모형이 적합한 것으로 보이며, 이모형을 최우추정법으로 추정한 결과가 표3.7에 주어져 있다.

모수추정치는 모두 유의하게 나타났으며 모형의 적합성을 알아보기 위한 포트맨토우 검정 결과 오차가 백색잡음과정을 따른다는 귀무가설을 기각하지 못하여 잔차의 SACF와 SPACF는 오차항이 백색잡음 계열임을 보여주므로 모형이 적합함을 알 수 있다.

따라서 대장암으로 인한 사망자수의 월별 시계열에 대한 승법계절ARIMA 모형은 (12)과 같다.

$$(1-B)(1-B^{12})y_t = (1-\underset{(0.06)}{0.79B})(1-\underset{(0.10)}{0.77B^{12}})a_t \ . \tag{12}$$

향후 2년간의 예측치는 표3.7과 같고, 그림3.3은 예측값과 예측오차의 시도표이다.

표3.6 대장암의 ARIMA(0,1,1)×(0,1,1)$_{12}$모형 적합 결과

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.78571	0.06331	12.41	<.0001	1
MA2,1	0.76855	0.10471	7.34	<.0001	12

Variance Estimate	383.8342
Std Error Estimate	19.59169
AIC	954.0066
SBC	959.3523
Number of Residuals	107

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.73	4	0.4435	-0.039	0.031	0.000	-0.072	0.010	0.157
12	9.19	10	0.5137	-0.016	0.008	-0.069	-0.124	0.080	-0.135
18	11.53	16	0.7759	0.010	0.069	-0.004	-0.052	0.095	-0.039
24	16.74	22	0.7776	0.093	0.065	-0.000	0.009	0.143	0.068
30	20.82	28	0.8328	-0.105	-0.039	-0.118	-0.025	-0.017	0.028
36	24.77	34	0.8765	0.046	-0.097	-0.030	0.108	-0.031	0.002
42	29.50	40	0.8888	0.087	0.082	-0.034	0.097	-0.028	0.045
48	32.63	46	0.9313	0.006	0.005	0.014	0.078	-0.020	0.097

Autocorrelation Plot of Residuals

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	383.834	1.00000	|********************|	0
1	-15.141364	-.03945	. *| .	0.096674
2	11.855744	0.03089	. |* .	0.096824
3	0.176173	0.00046	. | .	0.096916
4	-27.789476	-.07240	. *| .	0.096916
5	3.791991	0.00988	. | .	0.097420
6	60.297945	0.15709	. |***.	0.097430
7	-6.254677	-.01630	. | .	0.099769
8	2.962146	0.00772	. | .	0.099794
9	-26.408053	-.06880	. *| .	0.099799
10	-47.476727	-.12369	. **| .	0.100241
11	30.757255	0.08013	. |** .	0.101658
12	-51.870942	-.13514	.***| .	0.102246
13	3.823115	0.00996	. | .	0.103902
14	26.397503	0.06877	. |* .	0.103911
15	-1.441113	-.00375	. | .	0.104336
16	-19.993580	-.05209	. *| .	0.104337
17	36.633232	0.09544	. |** .	0.104580
18	-14.869716	-.03874	. *| .	0.105391
19	35.824424	0.09333	. |** .	0.105524
20	24.928196	0.06495	. |* .	0.106292
21	-0.172295	-.00045	. | .	0.106663
22	3.313616	0.00863	. | .	0.106663
23	54.887057	0.14300	. |***.	0.106669
24	25.979047	0.06768	. |* .	0.108446

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.03945	. *| .
2	0.02938	. |* .
3	0.00281	. | .
4	-0.07336	. *| .
5	0.00424	. | .
6	0.16346	. |***.
7	-0.00468	. | .
8	-0.01104	. | .
9	-0.06889	. *| .
10	-0.10893	. **| .
11	0.07565	. |** .
12	-0.15214	.***| .
13	-0.01236	. | .
14	0.07059	. |* .
15	0.03712	. |* .
16	-0.04317	. *| .
17	0.07120	. |* .
18	0.02208	. | .
19	0.07723	. |** .
20	0.04196	. |* .
21	-0.00323	. | .
22	-0.01839	. | .
23	0.17439	. |***.
24	0.09064	. |** .

표3.7 대장암 사망자수의 예측치 (ARIMA모형)

	2005년	2006년
1월	486.07	493.54
2월	475.92	486.90
3월	489.21	489.96
4월	475.25	488.80
5월	498.29	496.44
6월	480.10	487.21
7월	500.44	497.66
8월	485.98	499.85
9월	486.19	487.55
10월	506.14	498.54
11월	494.20	484.65
12월	478.28	480.34

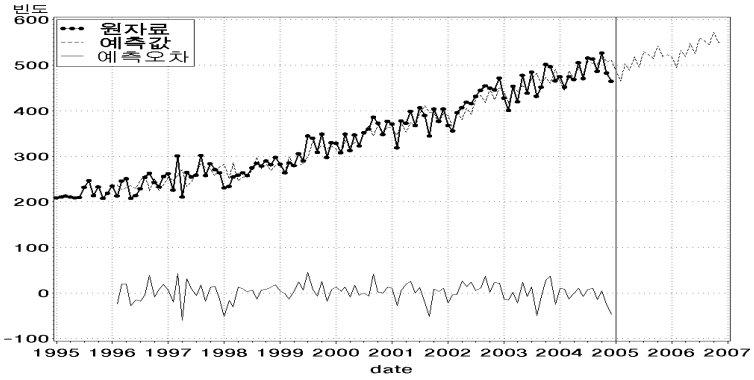


그림3.3 대장암의 사망자수와 예측 시도표(ARIMA)

3.4.2 지수평활법

대장암으로 인한 사망자수는 계절변동을 포함하고 있다. 이러한 계열에 대하여 예측을 시행할 때는 계절변동을 고려할 수 있는 계절 지수평활법(윈스터지수평활법)을 사용한다. 시계열의 계절적 변동폭이 일정한지 아닌지를 보고 승법 또는 가법 윈스터 방법을 사용하는데 대장암의 경우 계절적 변동폭이 일정하므로 가법 윈스터 방법을 사용하도록 하겠다.

지수평활법을 사용하기 위하여 SAS의 PROC FORECAST를 수행하여 적합하였다. 표3.8은 가법계절지수평활법의 적합통계량이고 표3.9는 예측결과이다. 표3.8을 살펴보면 $R^2 = 0.96$로 적합이 잘 되었음을 알 수 있다. 평활상수1~평활상수3은 0에 가까운 값을 가지므로 변화 정도가 서서히 이루어져 있다는 것을 의미한다. 계절요인의 효과에 대한 추정값을 살펴보면 환절기와 여름철에 사망자 수가 많아지는 것을 알 수 있다.

그림3.4는 예측값과 예측오차의 시도표이다. 예측오차들이 어떤 체계적인 상관관계가 존재한다면 예측오차에 예측의 정보가 남아 있어 예측값이 더 개선되어야 한다는 것을 의미하므로 이를 알아보도록 하자. 그림3.4에서 예측오차는 0에 관해 랜덤한 분포를 보이고 있고, 예측오차의 평균이 0인지에 대한 검정 결과 $t = 0.22 (p = 0.83)$으로 유의하지 않으므로 예측오차 간에 상관관계가 존재하지 않는다는 것을 알 수 있다. 또한 표3.10의 자기상관계수 역시 매우 작은 값을 가지므로 예측 오차들 간에 상관관계가 존재하지 않는 것으로 판단된다.

표3.8 가법계절지수평활법의 적합통계량

통계량	추정값	통계량	추정값
관측치(N)	120개	계절(1월)	-7.12
자유도(DF)	106	계절(2월)	-32.02
평활상수1	0.11	계절(3월)	3.12
평활상수2	0.11	계절(4월)	-11.62
평활상수3	0.25	계절(5월)	12.78
SST	1038440.80	계절(6월)	-12.68
SSE	43793.78	계절(7월)	18.30
MSE	413.15	계절(8월)	10.25
RMSE	20.33	계절(9월)	-0.68
MAPE	4.54	계절(10월)	23.36
MPE	-0.07	계절(11월)	-2.45
MAE	14.59	계절(12월)	-1.25
ME	0.38		
R^2	0.96		

표3.9 향후 2년간의 예측치(가법계절지수평활모형)

	2005년	2006년
1월	495.12	520.72
2월	472.36	497.96
3월	509.63	535.23
4월	497.02	522.63
5월	523.55	549.16
6월	500.23	525.84
7월	533.34	558.94
8월	527.43	553.03
9월	518.63	544.24
10월	544.81	570.41
11월	521.13	546.74
12월	524.46	550.07

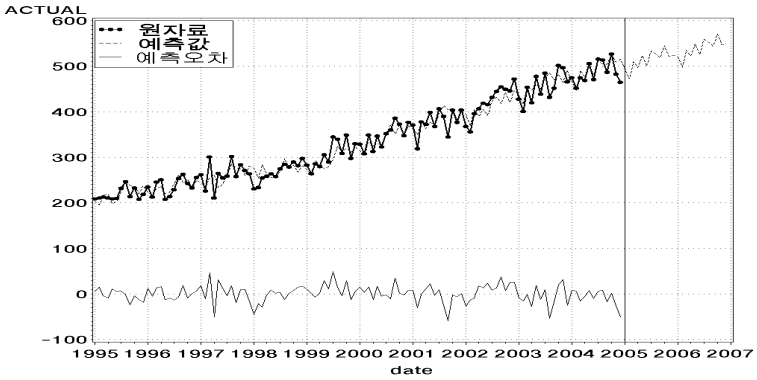


그림3.4 대장암의 사망자수와 예측 시도표 (가법계절지수평활모형)

표3.10 예측오차의 자기상관계수

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	364.800	1.00000	|**********|	0
1	20.138921	0.05521	|* . |	0.091287
2	47.273725	0.12959	|***. |	0.091565
3	16.719839	0.04583	|* . |	0.093081
4	-11.990898	-.03287	| * . |	0.093269
5	14.184569	0.03888	|* . |	0.093365
6	43.100298	0.11815	|** . |	0.093500
7	-8.138338	-.02231	| . |	0.094736
8	-3.822221	-.01048	| . |	0.094780
9	-44.793837	-.12279	| ** . |	0.094789
10	-54.518892	-.14945	|*** . |	0.096106
11	-6.373354	-.01747	| . |	0.098023
12	-62.697050	-.17187	|*** . |	0.098049

3.4.3 분해모형

대장암으로 인한 사망자수의 계절 변동이 일정하므로 가법 분해 모형을 사용하여 사망자수를 예측하여 보도록 하자.

그림3.2 시도표에서 알 수 있듯이 대장암 사망자수는 선형 추세를 가지고 있다. 그러나 추세성분(T_t)과 계절성분(S_t)은 서로 독립이 아니므로 추세성분

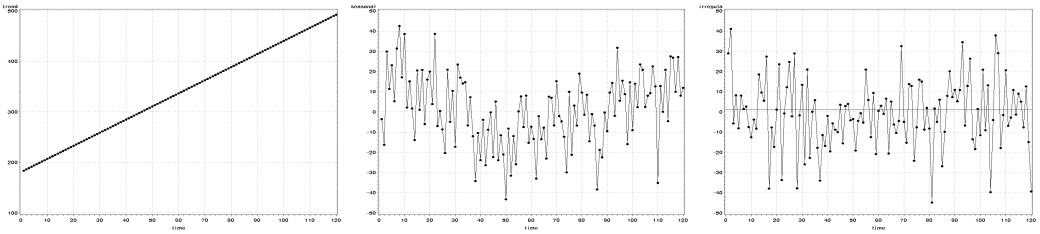
과 계절성분을 포함하는 추세모형을 이용하여 분해하는 것이 바람직하다. SAS의 PROC AUTOREG를 사용한 회귀분석 결과로부터 추세성분(T_t)및 계절성분(S_t)을 분해한 결과 잔차들 간에 상당히 큰 자기상관관계가 존재하여 ($DW=1.18(p<.001)$) 자기회귀오차모형을 적합 시켰다. 추정식은 (13)과 같다.

$$\begin{aligned}\hat{y_t} &= \underset{(0.06)}{2.58}t - \underset{(7.44)}{177.30}Jan - \underset{(7.47)}{154.02}Feb + \underset{(7.49)}{188.34}Mar - \underset{(7.51)}{169.86}Apr + \underset{(7.54)}{187.98}May \\ &+ \underset{(7.57)}{169.71}Jun + \underset{(7.59)}{195.93}Jul + \underset{(7.62)}{197.05}Aug + \underset{(7.64)}{182.37}Sep + \underset{(7.67)}{200.49}Oct \\ &+ \underset{(7.70)}{176.71}Nov + \underset{(7.72)}{184.94}Dec + \hat{a_t}\ , \\ \hat{a_t} &= \underset{(0.09)}{0.22}a_{t-1} + \underset{(0.09)}{0.23}a_{t-2} + \underset{(0.09)}{0.26}a_{t-6}\ . \end{aligned} \tag{13}$$

시계열의 변동요인을 개별성분(추세, 계절, 순환, 불규칙)으로 분해한 결과는 다음 그림3.5와 같다. 분해된 각 성분들을 개별적으로 예측한 후 다시 결합시켜서 예측한 예측치는 표3.11과 같고, 예측값과 예측오차의 시도표는 그림3.6과 같다.

불규칙 성분은 평균 0을 중심으로 랜덤한 패턴을 보이며 불규칙 성분의 자기상관계수도 거의 0에 가까워 성분추정이 잘 되었음을 확인시켜 준다.

<추세성분> <계절성분> <불규칙성분>



<불규칙성분의 자기상관계수>

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	285.123	1.00000	|**********|	0
1	-11.686769	-.04099	. *| . |	0.091287
2	-14.362045	-.05037	. *| . |	0.091440
3	19.511766	0.06843	.* . |	0.091671
4	-1.152876	-.00404	. . |	0.092096
5	24.692113	0.08660	.** . |	0.092097
6	2.388448	0.00838	. . |	0.092774
7	7.695645	0.02699	.* . |	0.092780
8	-6.866493	-.02408	. . |	0.092845
9	-16.823820	-.05901	. *| . |	0.092897
10	-14.353627	-.05034	. *| . |	0.093209
11	34.493129	0.12098	.** . |	0.093435
12	-31.035017	-.10885	. **| . |	0.094732

그림3.5 분해법에 의해 추정된 추세·계절·불규칙성분

표3.11 향후 2년간의 예측치(분해모형)

	2005년	2006년
1월	483.73	513.37
2월	462.84	492.47
3월	499.86	529.50
4월	484.87	514.51
5월	505.07	534.71
6월	489.02	518.65
7월	518.32	547.96
8월	522.51	552.15
9월	509.58	539.21
10월	531.17	560.81
11월	509.53	539.17
12월	521.42	551.06

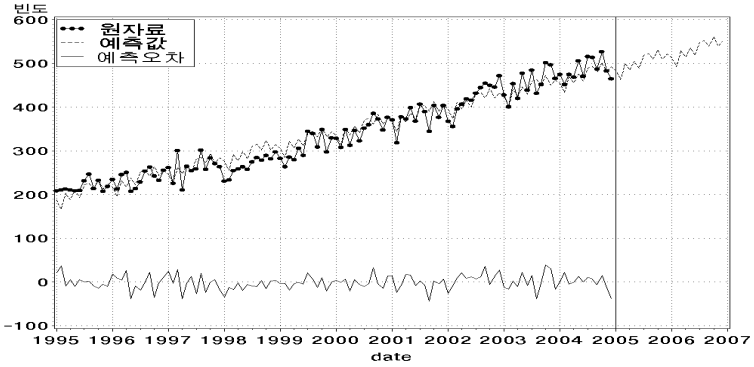


그림3.6 대장암의 사망자수와 예측 시도표 (분해법)

3.4.4 대장암 시계열 모형 적합에 대한 결론

대장암으로 인한 사망자수를 3가지 분석법으로 향후 2년간의 사망자수를 예측하여 보았다. 어느 모형과 분석법을 자료에 적용하는 것이 타당한지 알아보기 위하여 예측력의 측면에서 살펴보도록 하자.

3가지 모형 중 최적 모형을 선택하는 기준으로 제곱근평균제곱오차(RMSE), 절대오차퍼센트평균(MAPE)를 사용하도록 하겠다.

표3.12 예측방법별 RMSE, MAPE 비교

예측방법	RMSE	MAPE
승법계절 ARIMA	19.46	4.43
가법계절지수평활	20.33	4.54
분해법	20.72	5.48

표3.12는 각 예측방법별 RMSE와 MAPE통계량 값이다. 모형적합 통계량인

RMSE와 MAPE는 0에 가까울수록 모형이 잘 적합 되었다고 말하는데 세 가지 예측 방법 중에서 승법계절 ARIMA의 적합통계량이 가장 작으므로 예측의 측면에서 ARIMA방법에 의한 예측이 지수평활법이나 분해법보다 더 나은 것으로 생각된다.

그림3.3, 그림3.4, 그림3.6을 보면 각 예측방법별 추이를 볼 수 있다. 3가지 예측방법에 의하여 대장암으로 인한 사망자수는 꾸준히 증가할 것으로 보인다. 하지만 ARIMA방법에 의한 예측은 예측 시점이 커질수록 증가 추이가 줄어드는 ARIMA모형의 성질에 의하여 단기예측에 효과적으로 사용되며 평활법과 분해법에 의한 예측은 시계열의 구성요소가 시간에 의존하여 천천히 움직이는 경우 적당하며 주로 중기 예측 이상에 주로 사용된다(박유성, 김기환; 2001).

제4장 결 론

본 논문은 우리나라 사망원인 자료를 이용하여 사망순위를 부여하고, 성, 연령, 결혼상태, 교육수준에 따른 빈도분석, 교차분석 및 수량화 방법을 이용한 기초분석을 실시하였다.

사망원인이 4가지 설명변량(성, 연령, 결혼상태, 교육수준)에 따라 유의한 연관성을 보였다. 성별로는 악성신생물, 운수사고, 간질환, 자살, 고혈압성질환, 호흡기 결핵이 큰 차이를 보였으며, 이 중 고혈압성 질환을 제외한 나머지는 남성의 비율이 더 높았다. 연령별로는 20대 이하 연령대에서는 사고사나 선천적인 사인으로 인한 사망자가 많았고, 그 이후 연령대에서는 악성신생물이나 뇌혈관 질환으로 인한 사망자가 많았다. 결혼 상태에 따라서는 미혼인 경우 운수사고나 자살의 사망이 가장 많았고, 배우자가 있거나 이혼, 사별은 질환에 의한 사망이 많았다. 이들의 차이는 사망자 연령과 관계가 있을 것으로 생각된다.

수량화 방법Ⅱ를 이용한 10대 사망원인에 대한 연관성분석 결과, 연령이 높아질수록, 여성일수록, 주소지가 대도시 일수록 질환에 의한 사망빈도가 높았다.

우리나라 사인 순위 선정시 사용되는 56항목의 각 사인별 월별 사망자수를 시계열 분석방법으로 예측하였다. 사망자수 자료에 ARIMA모형을 적합하여 예측한 결과 운수사고, 간질환, 호흡기 결핵 등은 감소하는 추세에 있으며 악성신생물, 당뇨병, 자살 등은 증가하는 추세를 보였다. 특히 우리나라 사망원인 1위인 악성신생물을 26가지 장기별로 ARIMA모형을 적합하여 예측한 결과 13개 장기가 ARIMA(0,1,1)모형으로 적합 되었다. 폐, 대장, 췌장, 담낭 등은 증가 추세에 있고, 후두만 감소추세에 있었다.

한국인의 사망률을 낮추기 위해서는 각종 사고로 인한 희생을 최소화하는 노력이 필요하며 향후 증가 추이를 보이는 암 또는 질환에 대하여 조기검진과 같은 체계적인 시스템과 홍보를 통하여 예방하는 것이 중요하다. 현재 시행하고 있는 암 검진과 같은 시스템을 앞으로 큰 증가추이를 보일 심·뇌혈관 질환에 대하여서도 적용하여 국가적인 관리와 지원이 필요하다고 생각된다.

본 논문에서 제시한 기초분석 및 예측모형은 의료 정책 결정 혹은 의학 연구의 기초자료로 활용 가능하다. 향후 연구에서는 다변량시계열 분석 혹은 자기회귀시차분포모형(Autoregressive Distributed Lag Model: ADL모형)등을 이용하여 개선된 예측결과 얻는 연구·분석이 필요할 것이라고 생각된다.

참 고 문 헌

[1] 박유성, 김기환 (2002). SAS/ETS를 이용한 시계열자료분석Ⅰ. 자유아카데미. 서울.

[2] 이동우, 김일순 (1997). 사망력 지표의 개발 및 측정-사망신고 자료를 중심으로, 한국의 보건문제와 대책(Ⅱ). 415~452.

[3] 이종협, 최기헌 (1994). SAS/ETS를 이용한 시계열 분석과 그 응용. 자유아카데미. 서울.

[4] 정동빈, 원태연 (2005). SPSS를 이용한 시계열자료와 단순화분석. SPSS아카데미. 서울.

[5] 통계청 (1995-2004). 사망원인통계.

[6] 통계청 (2004). 한국표준질병·사인분류 제2권 지침서. 8-130.

[7] 허명회 (1998). 수량화방법Ⅰ·Ⅱ·Ⅲ·Ⅳ. 자유아카데미. 서울.

[8] 林知己夫, 樋口伊佐夫, 駒澤勉 (1970). 情報處理と統計數理. 産業圖書.

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

[10] Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes, *Annals of the Institute of Statistical Mathematics.* 26. 363-387.

[11] Box, G. E. P. and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-intergrated moving average time series models. *Journal of American Statistical Association.* 65. 1509-1526.

[12] SAS/ETS User's Guide 9.1 (2006). SAS Institute Inc.

[13] SAS/STAT User's Guide 9.1 (2006). SAS Institute Inc.

[14] Schwarts, G. (1978). Estimating the dimension of a model, *Annal of Statistics*. 6. 461-464.

ABSTRACT

Time Series Analysis on Major Diseases of Korean Mortality Data

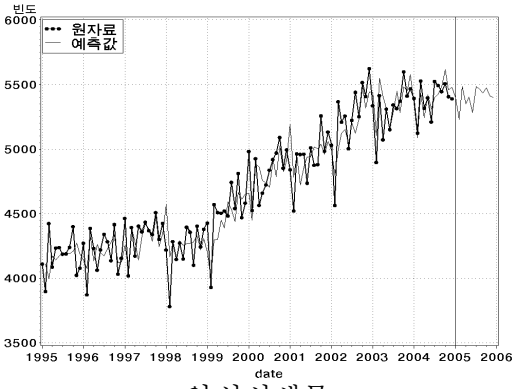
Young-eun Kim
Department of Statistics
The Graduate School
Sungshin Women's University

In case of Korea, since there is rapid development and industrialization, the structure of the death cause rapidly changed. So, it's important to analyze and predict the time series trend of the death cause at the first stage of the aging society.

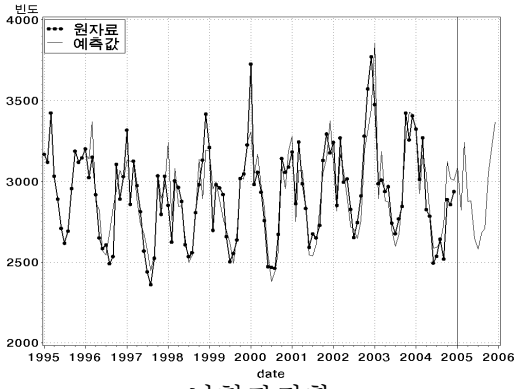
This thesis investigates the overall trend in the main cause of death using the frequency analysis, cross-tabulation analysis and quantification method Ⅱ on major diseases' Korean mortality data from 1995 to 2004. It also conducts the ARIMA model fitting for the individual cause of death of 56 items which are selected as the high rank of the cause of death and forecasting them future values.

This thesis carries out forecasting future values of the death of 26 cancers using the ARIMA models. It also compares the forecasting ability among the ARIMA model, the smoothing method, the decomposition model fitted to the death of the colon cancer.

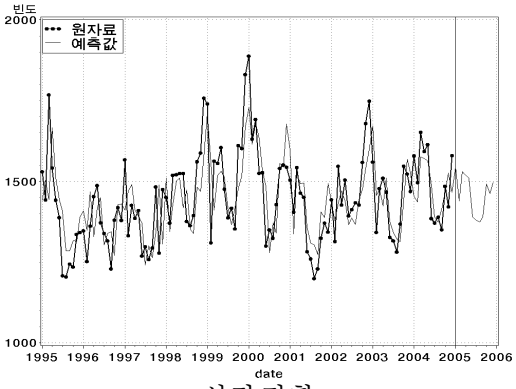
부록A. 56개 항목 사인별 예측 시도표



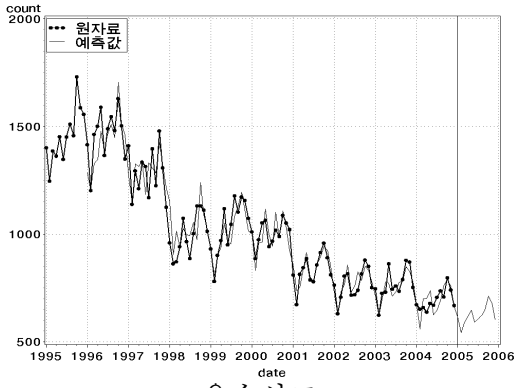
<악성신생물>



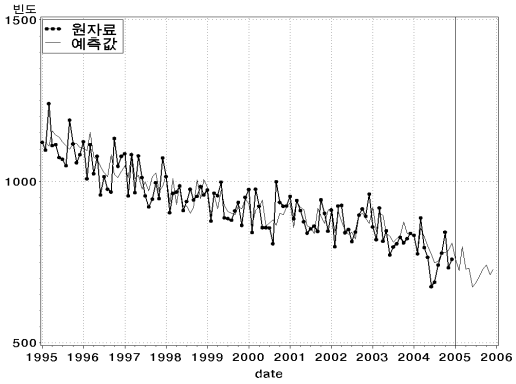
<뇌혈관질환>



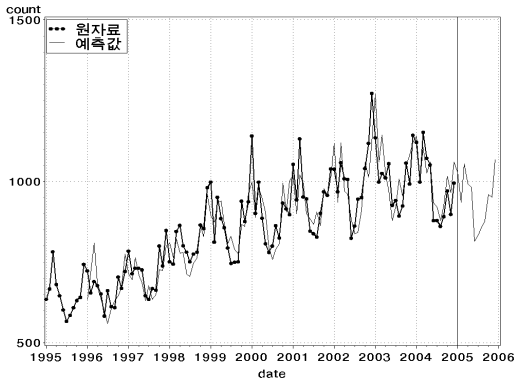
<심장질환>



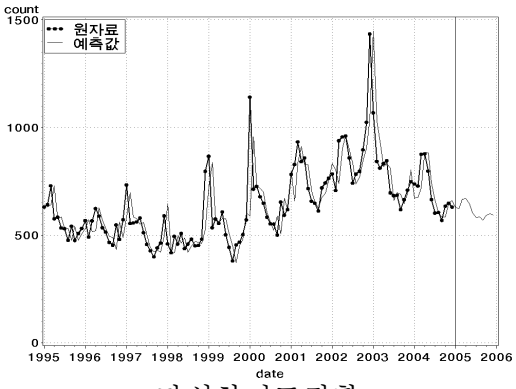
<운수사고>



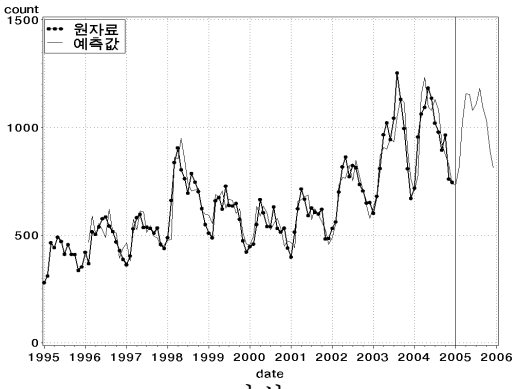
<간질환>



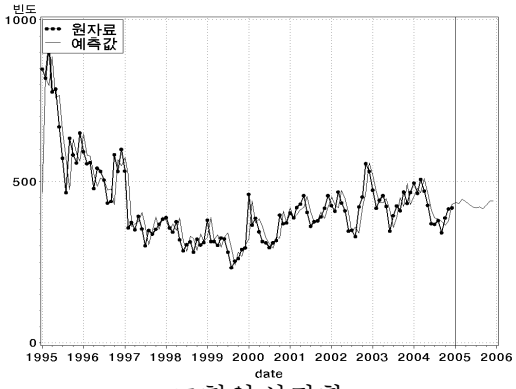
<당뇨병>



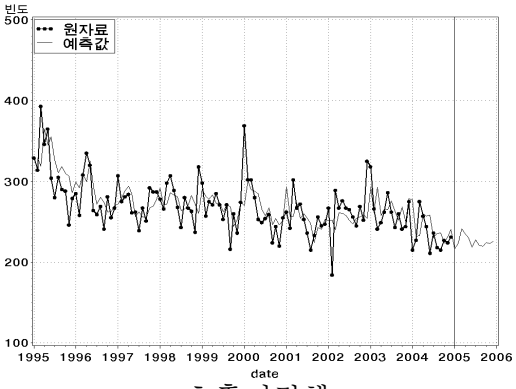
<만성하기도질환>



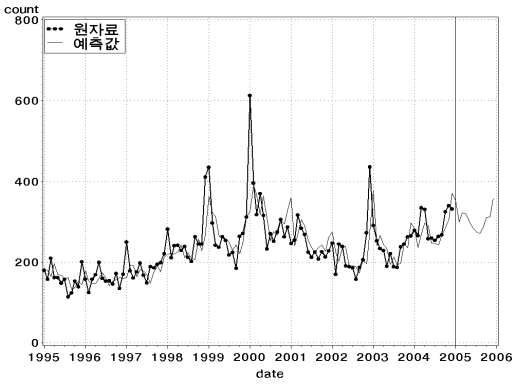
<자살>



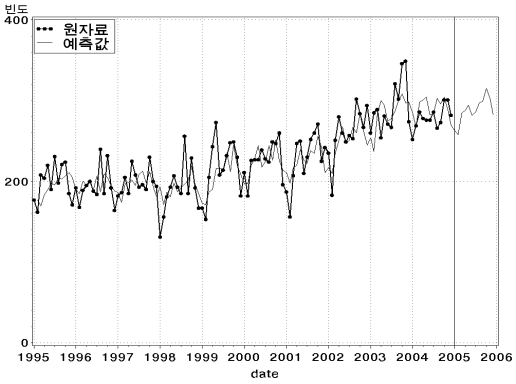
<고혈압성질환>



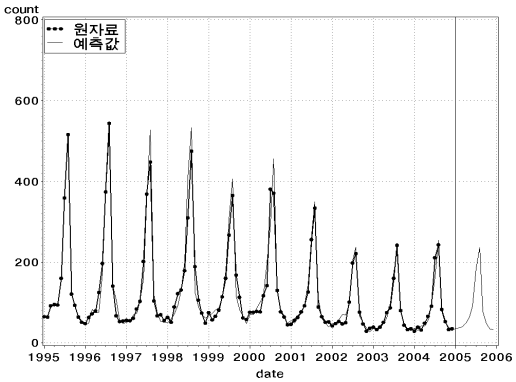
<호흡기결핵>



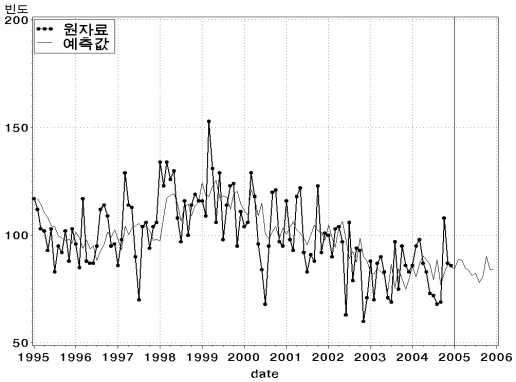
<폐렴>



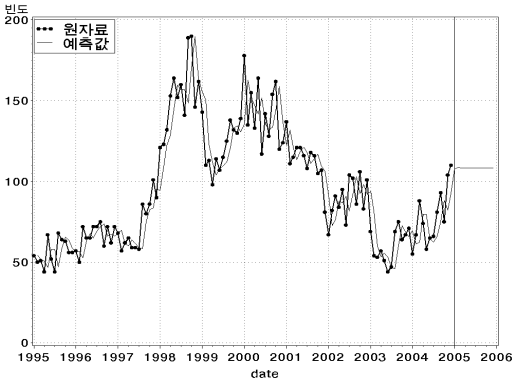
<추락>



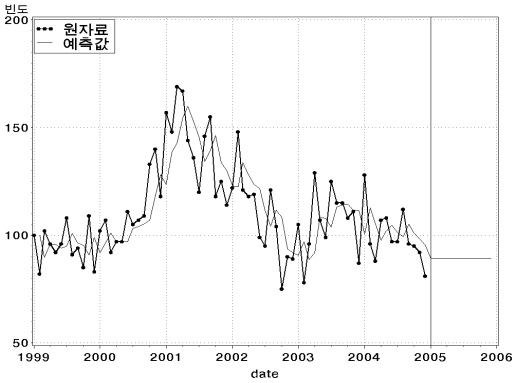
<익사>



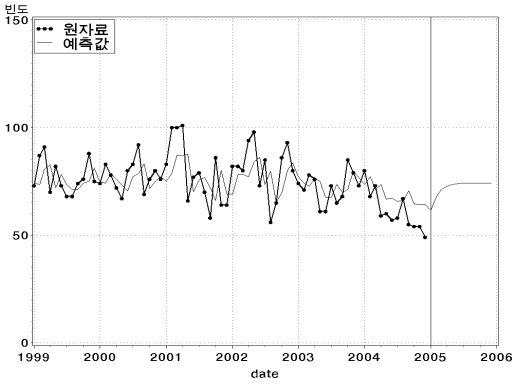
<정신활성물질 사용에 의한 정신및 행동장애>



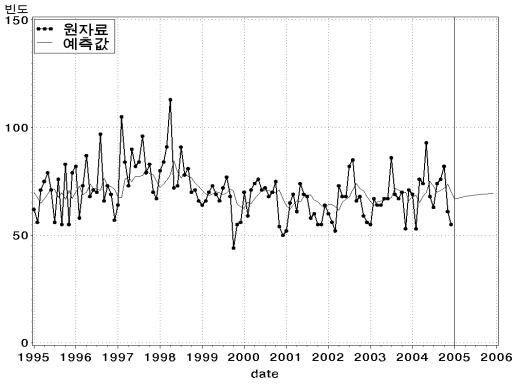
<패혈증>



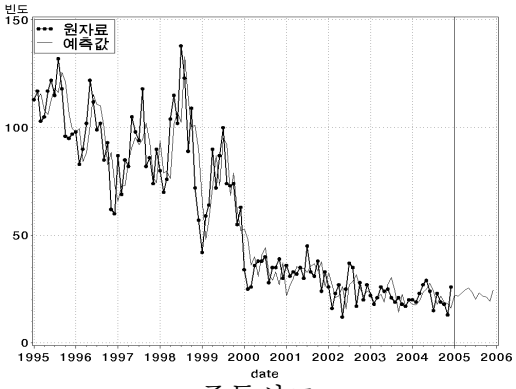
<출생전후기질환>



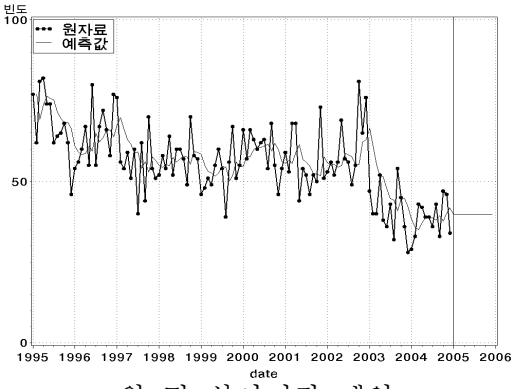
<선천기형>



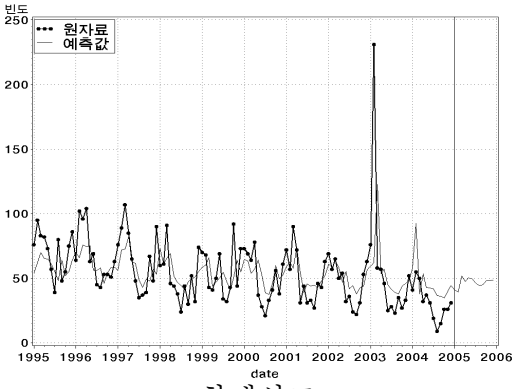
<타살>



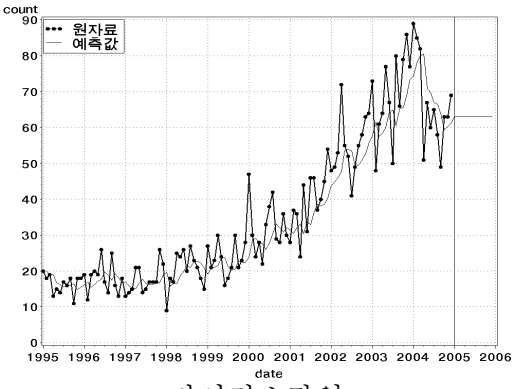
<중독사고>



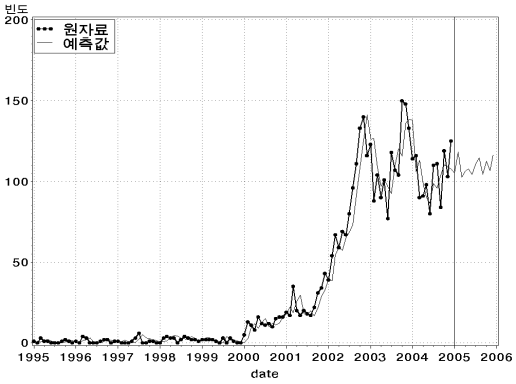
<위 및 십이지장 궤양>



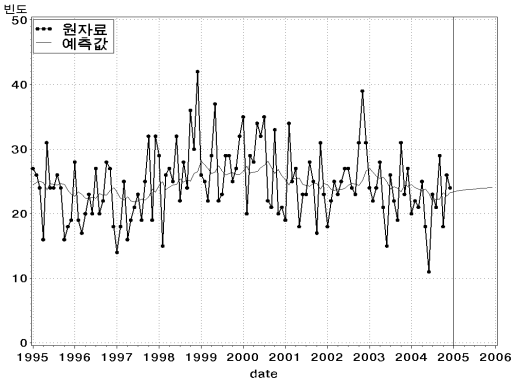
<화재사고>



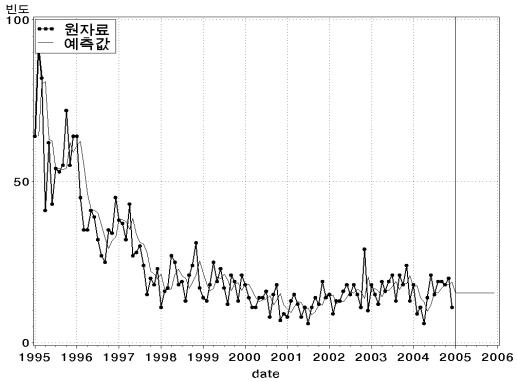
<바이러스간염>



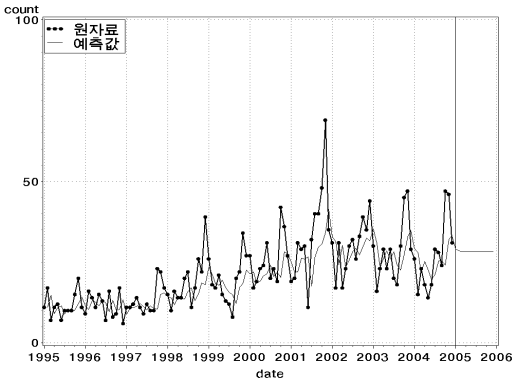
<알쯔하이머병>



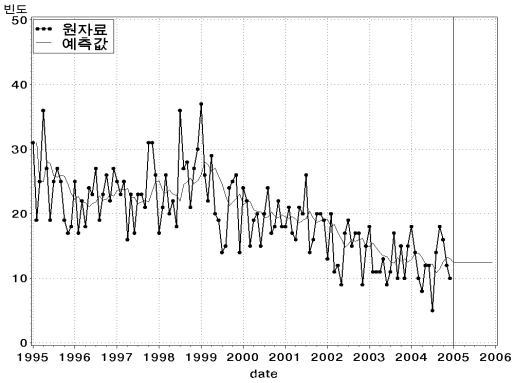
<빈혈>



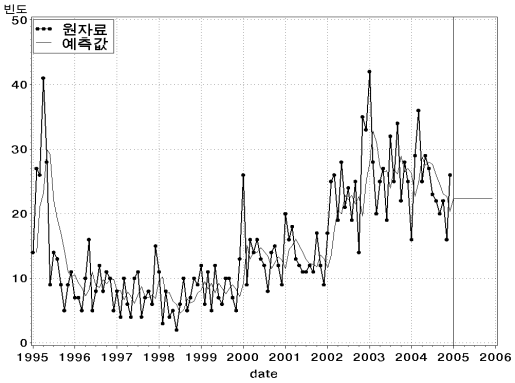
<죽상경화증>



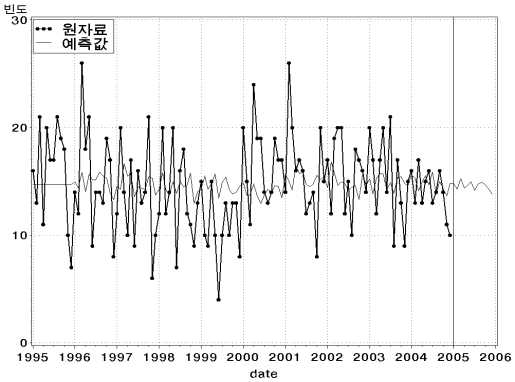
<나머지 특정 감염성 및 기생충성 질환>



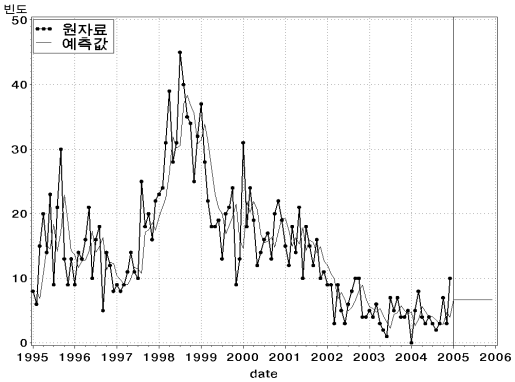
<사구체질환 및 세뇨관-사이질성 질환>



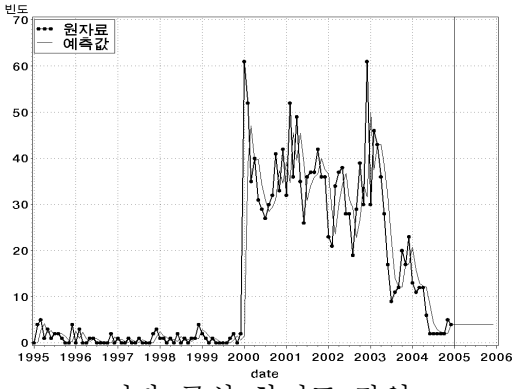
<급성 류마티스열 및 만성 류마티스 심장질환>



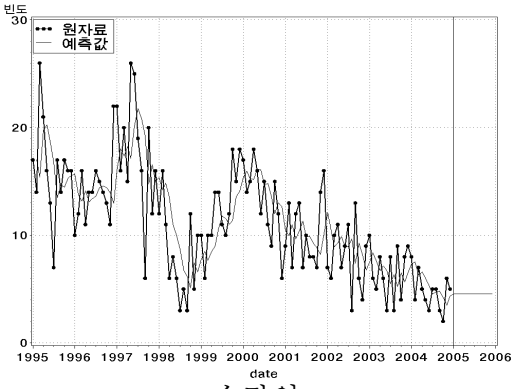
<기타결핵>



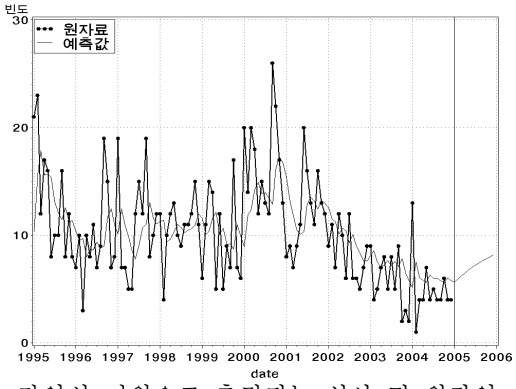
<영양실조>



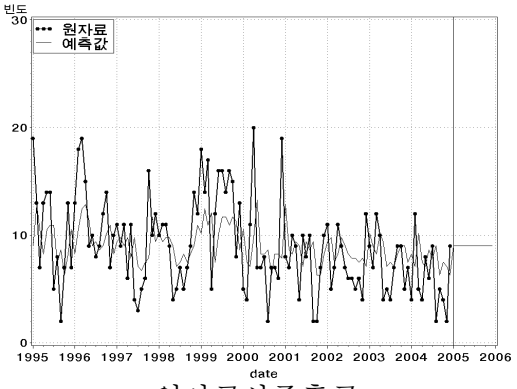
<기타 급성 하기도 감염>



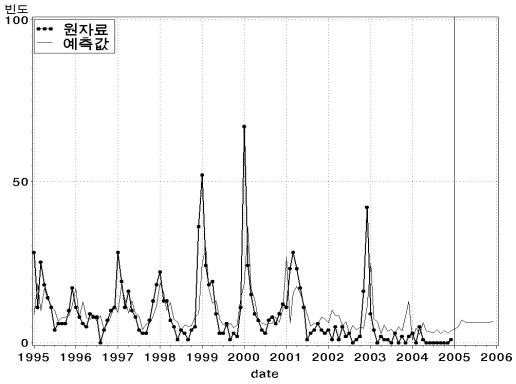
<수막염>



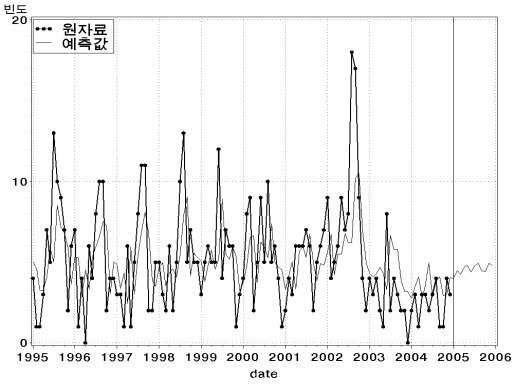
<감염성 기원으로 추정되는 설사 및 위장염>



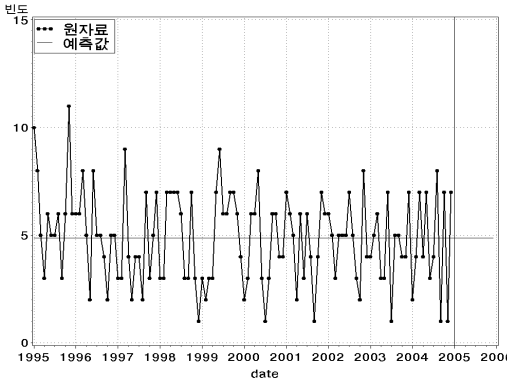
<영아급사증후군>



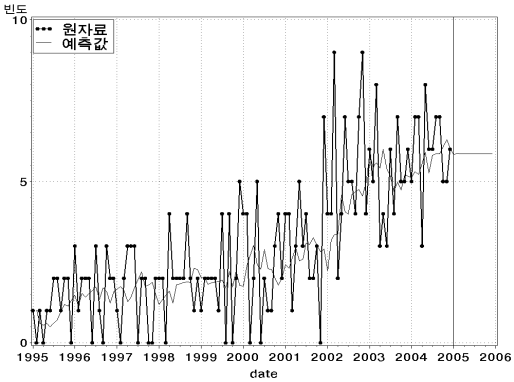
<인플루엔자>



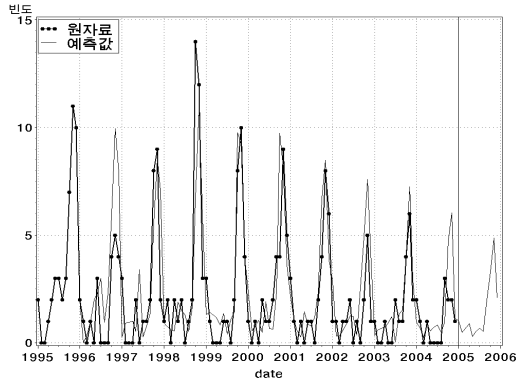
<기타 창자 감염성 질환>



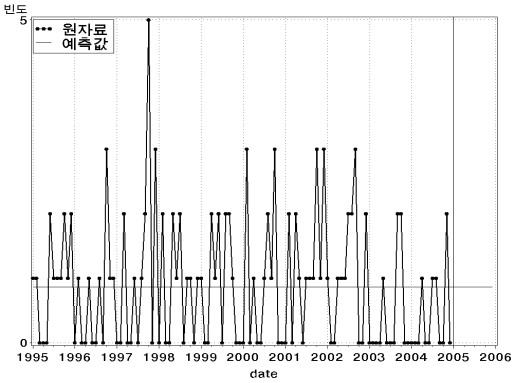
<기타 직접 산과적 사망>



<인체 면역결핍 바이러스병>

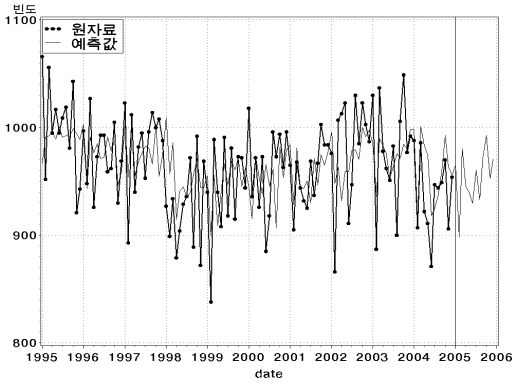


<기타 절지동물 매개의 바이러스열 및 바이러스 출혈열>

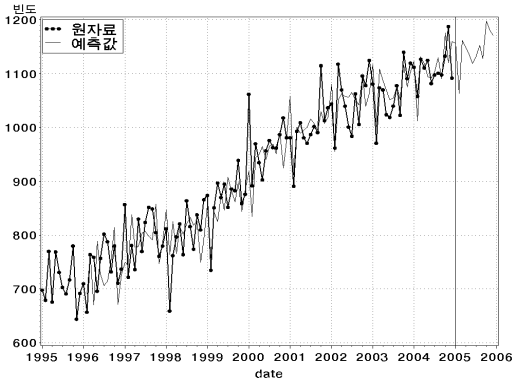


<파상풍>

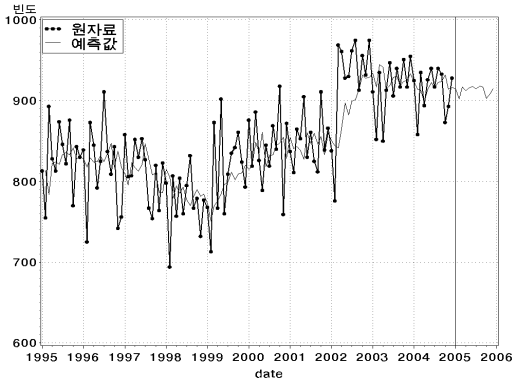
부록B. 26가지 악성신생물별 예측 시도표



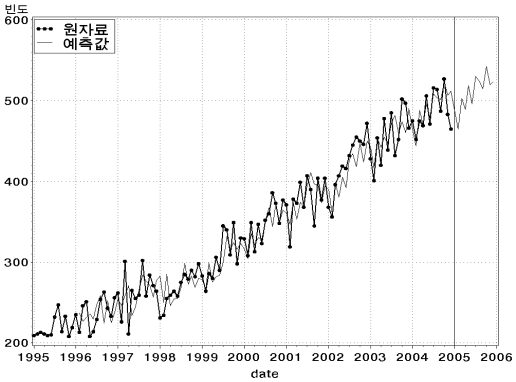
<위>



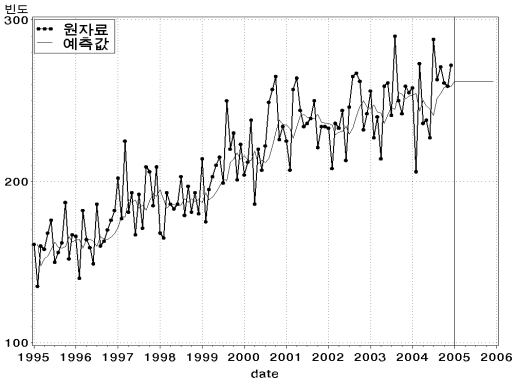
<폐>



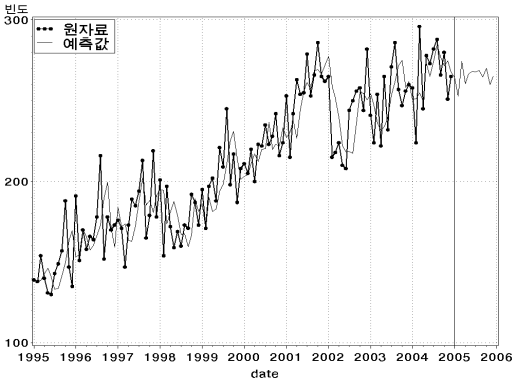
<간>



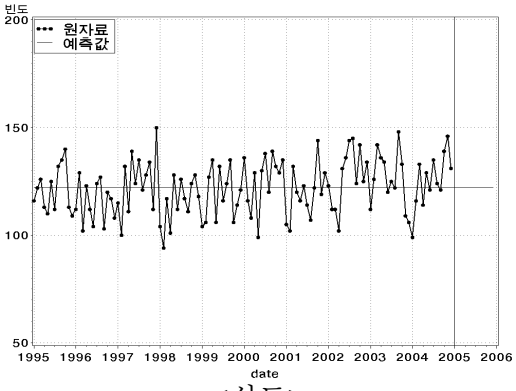
<대장>



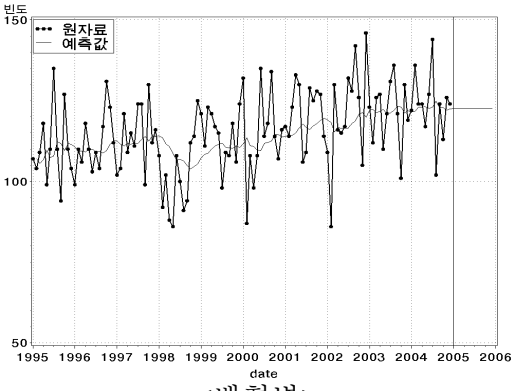
<췌장>



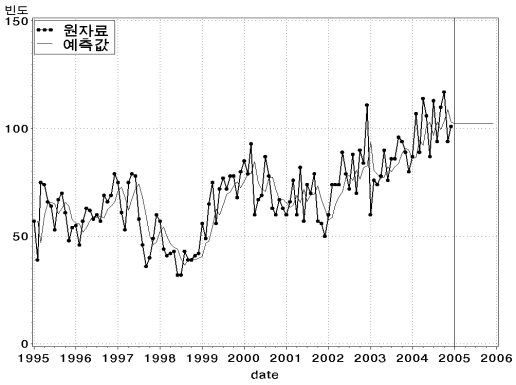
<담낭>



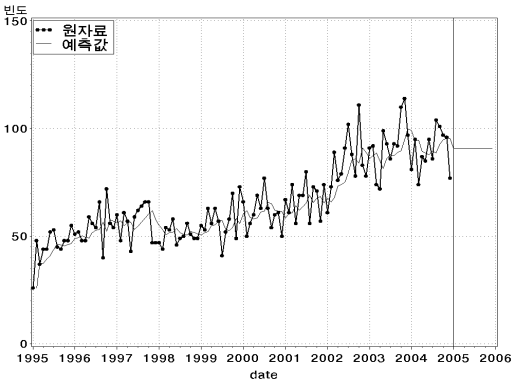
<식도>



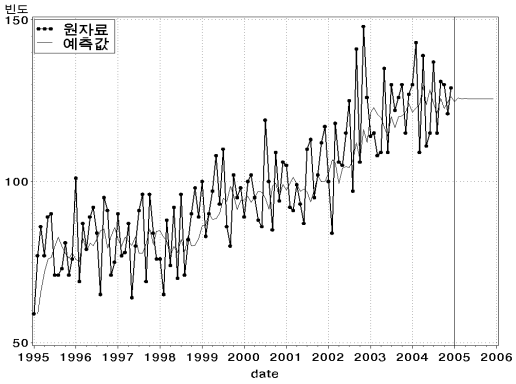
<백혈병>



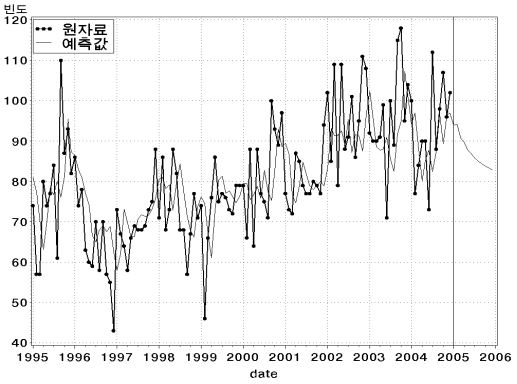
<비호지킨림프종>



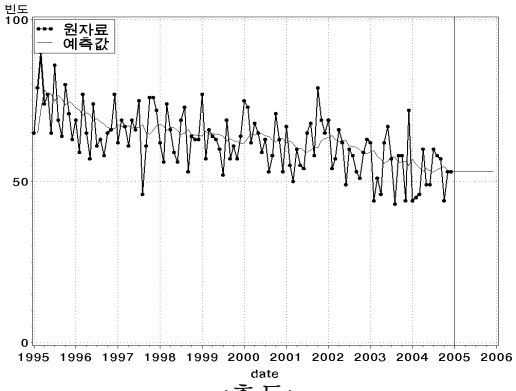
<자궁경부>



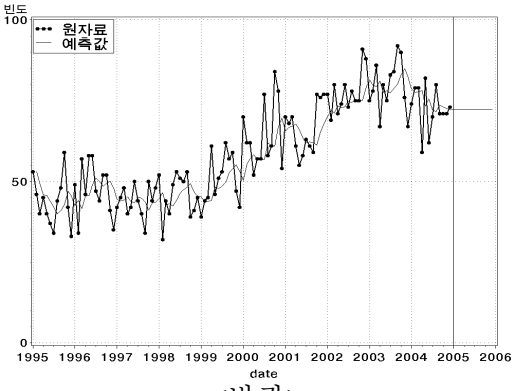
<유방>



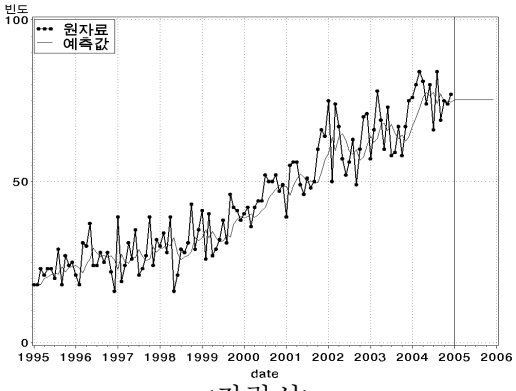
<뇌 및 중추신경계>



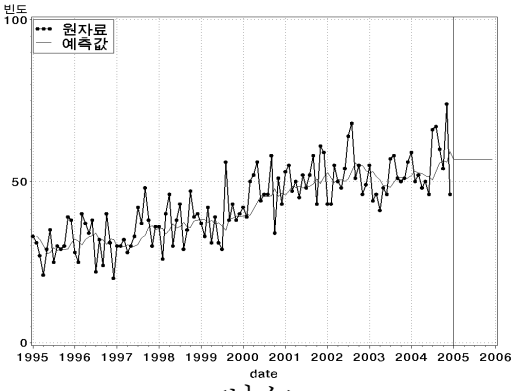
<후두>



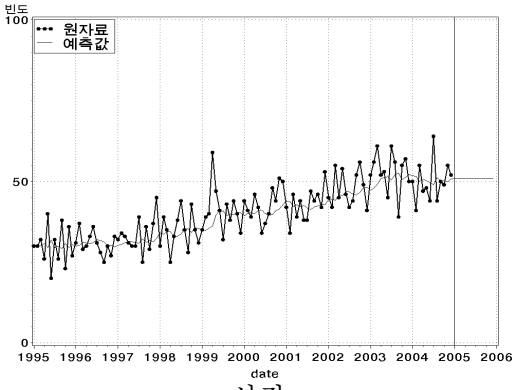
<방광>



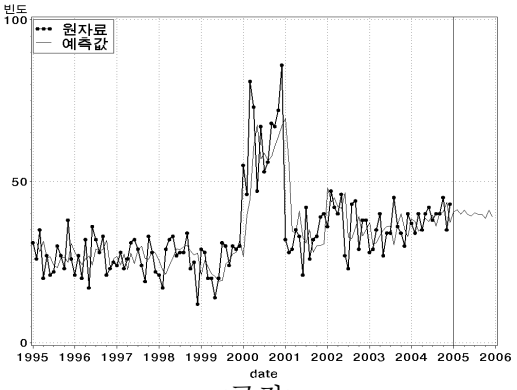
<전립선>



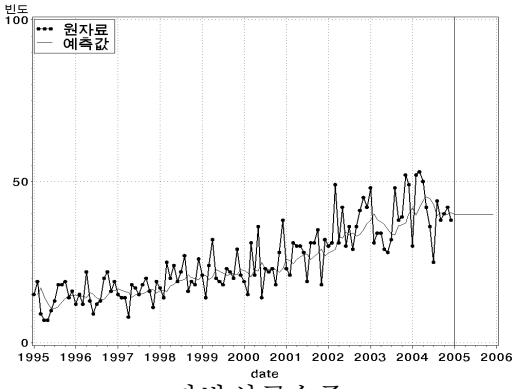
<난소>



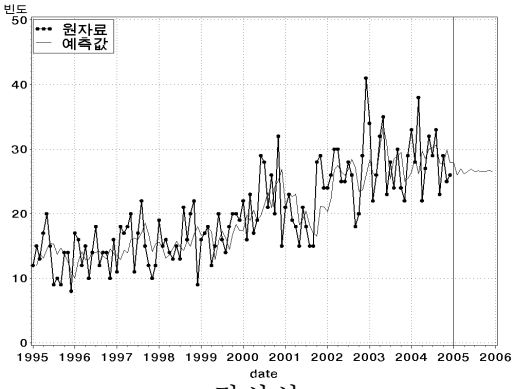
<신장>



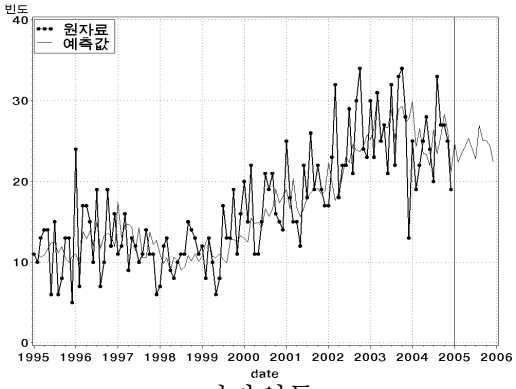
<구강>



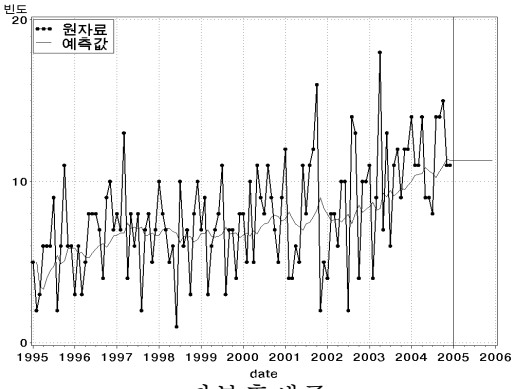
<다발성골수증>



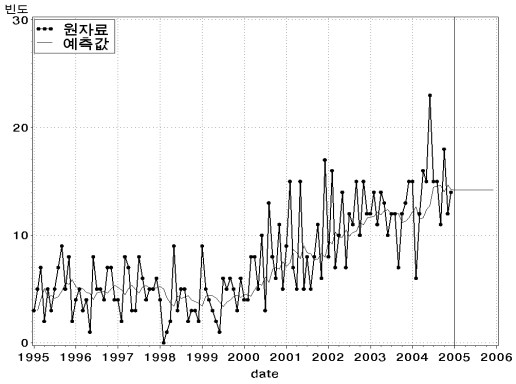
<갑상선>



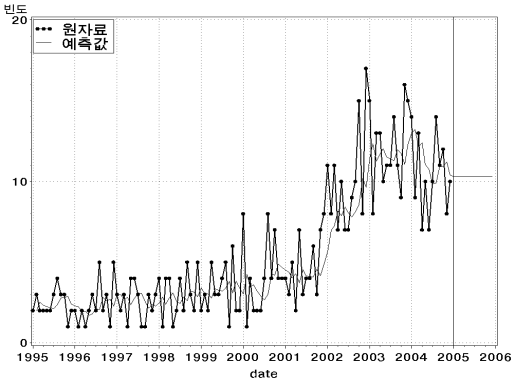
<기타인두>



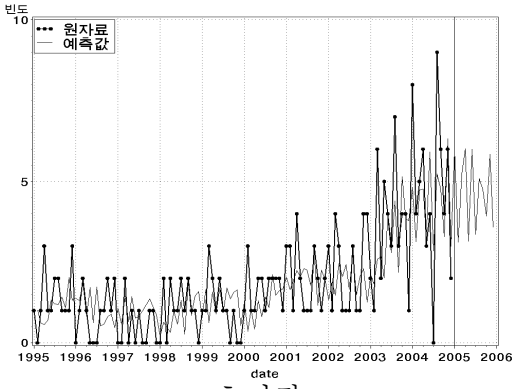
<피부흑색종>



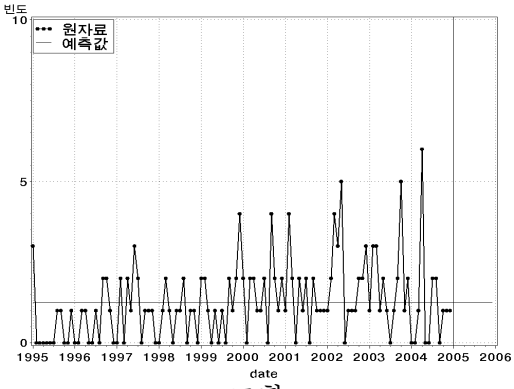
<비인두>



<자궁체부>


빈도
10
5
0
원자료
예측값
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
date

<호지킨>


빈도
10
5
0
원자료
예측값
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
date

<고환>

감 사 의 글

어느덧 시간이 흘러 성신교정에서 여섯 번째 겨울을 맞이하였습니다. 힘들어 포기하고 싶었던 순간들도 있었지만 항상 용기를 주시고 따뜻한 격려로 다시 일어날 수 있는 힘을 주셨던 송일성 교수님, 이해용 교수님, 이우선 교수님, 이종협 교수님, 이성건 교수님께 먼저 감사의 뜻을 올립니다.

학부1학년 때부터 대학원까지 옆에서 함께 걸어준 내 친구 희라, 주현이와 논문 쓰는 내내 힘내라고 응원해준 착한 후배 인경, 지윤, 희원이, 바쁜 회사 생활에도 불구하고 많은 도움을 주었던 여러 선배님들, 대학원 생활을 함께한 주영언니와 가영언니께 감사함을 전합니다. 눈빛만 보아도 통하는 소꿉친구 윤경, 미주, 미경이, 즐거운 대학생활을 함께 만들었던 지영, 윤정, 미은, 혜정, 혜원이 에게도 감사의 뜻을 전합니다.

툴툴거리는 동생 힘내라고 함께 소주잔을 기울여주던 필상오빠, 준호오빠, 소희언니에게 감사하며, 동아리 풀과비 사람들에게도 감사드립니다.

어리고 철없는 딸의 투정을 묵묵히 받아준 사랑하는 우리 엄마, 멀리 있지만 그 누구보다 나를 사랑해 주는 우리 아빠, 말은 하지 않아도 동생의 힘찬 도약을 바라는 오빠에게 감사의 마음과 사랑을 전하고 싶습니다.

감사하다는 말로는 차마 다 표현하지 못할 가르침을 주시고, 더 큰 세상을 향해 날아갈 수 있도록 날개를 달아주신 이종협 교수님께 온 마음과 정성을 다하여 감사드립니다.

그 외에 도움을 주신 많은 분들께 감사드립니다.

받은 만큼 돌려줄 수 있는 사람이 되도록 열심히 노력하겠습니다.