이 종 협 교수지도

석사학위청구논문

# 주요 질환 사망자료에 대한 시계열 분석

2006

성신여자대학교 대학원

통 계 학 과

김 향 선

# 주요 질환 사망자료에 대한 시계열 분석

이 종 협 교수지도

이 논문을 석사학위논문으로 제출함

2005년 11월

성신여자대학교 대학원

통 계 학 과

김 향 선

# 인 준 서

김향선의 석사학위 논문으로 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

# 논 문 개 요

사망은 정치, 경제, 사회 및 문화의 다양한 요인들을 포함한다. OECD회원국인 우리나라의 최근 주요 사망원인이 전염성 질환보다 비전염성질환으로 인한 사망이 높아지고 있다.

본 논문에서는 1992년 1월부터 2002년 12월까지의 미국 사망 자료를 이용하여 10대 사망원인에 대한 전반적인 경향을 빈도분석 및 수량화방법Ⅱ를 이용한 탐색적 방법으로 분석한다. 사망원인 중 성인병 질환인 허혈성심장병, 뇌혈관질환, 고혈압 및 당뇨병에 대해 회귀모형 및 승법계절 시계열 모형을 적합시킨 후 최적 예측모형을 선택하고 각 질환에 대한 사망빈도를 예측한다. 미국의 사망원인 패턴을 통해 우리나라의 주요 사망원인에 대한 향후 변화추이를 진단한다.

# 목 차

# 제1장 서론

사망은 개인의 생물학적 특성 및 행위의 결과인 동시에 사회, 문화, 정치, 경제적 요인들의 총체적 소산이다. 따라서 사망구조를 파악하는 것은 사회, 문화 및 보건에 걸쳐 다양한 분야를 이해하는 것이며 이것은 그 집단의 삶의 질을 향상시킬 수 있는 지표가 될 수 있다.

선진국의 경우 생활수준 향상과 의학 및 예방의학의 발전으로 국민보건이 급격히 향상되어 이환률이나 사망률, 그리고 질병의 양상에 큰 변화를 초래하였다. 즉, 경제발전으로 복지시설이 확대되어 환경이 개선되고 있어 과거에 많았던 각종 전염병은 감소되고 있는 반면 노년인구의 증가와 더불어 비전염성질환 즉, 만성 퇴행성질환이 증가하고 있다.(박홍현,2002) 우리나라 통계청의 2004년 사망통계 분석결과를 살펴보면 10년 전에 비하여 특정 감염성 및 기생충성, 순환기계통, 소화기계통 질환, 사망의 외인(사고사) 사망률은 감소하였으며 신생물, 내분비·영양·대사질환, 호흡기계통질환 사망률은 증가하고 있다. 한국의 경우 빠른 경제 성장과 OECD에 가입하는 등의 활발한 외교 활동으로 선진국 반열에 들어서고 있다. 우리나라 사망원인 역시 전염성 질환보다 비전염성 질환의 사망률이 높게 나타나고 있으며 앞으로 우리나라에서 높은 사망률이 예상되는 질환들 또한 선진국과 비슷한 패턴을 보일 것으로 예상되어진다. 따라서 본 논문에서는 미국의 사망자료를 토대로 10대 주요 사망원인에 대한 기초분석을 실시하여 사망원인의 일반적인 특성을 살펴보고 향후 우리나라 사망원인의 패턴을 파악하고자 한다. 또한 10대 사망원인 가운데 우리사회에서 점차 높은 사망률을 보이고 있는 성인병 질환인 허혈성심장병, 뇌혈관질환, 고혈압, 당뇨병에 대한 심도있는 분석을 통해 사망패턴을 예측함으로써 우리나라의 성인병으로 인한 사망특

성을 파악하는데 도움이 되고자한다.

본 논문의 구성은 다음과 같다. 제2장에서는 미국 사망원인에 대한 기초자료를 설명하고, 보건 및 의학계에서 일반적으로 사용하는 설명변수인 성별, 연령, 결혼상태 및 교육수준에 대하여 빈도분석을 실시한다. 또한 수량화 방법을 이용한 탐색적 다변량 자료분석을 수행한다. 제3장에서는 관심이 높은 성인병 질환인 허혈성 심장병, 뇌혈관질환, 고혈압, 당뇨병에 대하여 1992년 1월부터 2002년 12월까지의 미국 시계열 자료를 이용하여 선형계절 추세모형, 오차항이 AR과정을 따르는 회귀모형, 승법계절 ARIMA모형을 적합시킨다. 최적모형을 선택한 후 향후 성인병 질환에 대한 사망빈도를 예측한다. 마지막으로 제4장에서 본 연구의 결론을 맺는다.

# 제2장 기초분석

## 2.1 분석 데이터

미국의 사망원인에 대한 기초분석을 실시하기 위하여 NBER(National Bureau of Economic Research)의 Vital Statistics NCHS's Multiple Cause of Death Data 중 Mortality Data를 이용하였다. Mortality Data는 1992년 1월부터 2002년 12월까지의 자료로 구성되어 있다.

NBER에서는 사망자 발생시 작성된 사망증명서를 기준으로 Mortality Data를 수집하였는데 사망증명서에 적힌 항목들을 요약하면 다음과 같다.

<표 1> 사망증명서에 적힌 항목들

| 항목 | 하위 항목 |
|---|---|
| 1. 일반적인 사항 | a. 사망연도 / 사망월 / 사망요일<br>b. 성별(남 /여)<br>c. 연령(10대 이하 /20대 /30대 /40대 /50대 /60대/ 70대 /80대 이상)<br>d. 인종(백인 /황인 /흑인 혹은 14개의 분류기준)<br>e. 결혼상태(미혼 /기혼 /미망인 /이혼)<br>f. 직업<br>g. 교육수준(8년미만/9년-11년/12년/13-15년/16년이상) |
| 2. 사망지역정보 | a. 사망장소의 주와 군에 대한 정보<br>b. 사망발생지역을 뉴욕의 다섯개 행정구, 시카고와 기타로 구별함<br>c. 1980년도 인구조사를 기준으로 2.a의 해당인구 |
| 3. 거주지 | a. 거주 장소의 주와 군에 대한 정보<br>b. 1980년도 인구조사를 기준으로 3.a의 해당인구 |
| 4. 사망원인 | a. 사망분류기준(ICD-9/ICD-10)정보<br>b. 각 분류기준에 따른 사망원인 항목<br>c. 직접적으로 작용한 사인과 간접적으로 작용한 사인에 대한 항목 |

<표 1>에서 보듯이 각 항목에 대한 다양한 하위항목들이 존재하는데 본 논문에서 사용하는 설명변수는 의학 및 보건통계 분야에서 가장 보편적으로

사용되고 있고 우리나라 사망진단서에 기록되는 내용과 공통되는 부분인 항목1(일반적인 사항)에 있는 사망연도, 사망월, 성별, 결혼상태, 교육수준이다.

항목4에 기록된 사망원인을 분류하는 기준으로 ICD code를 사용하고 있는데 ICD란 국제질병분류(International Classification of Diseases)의 약자로써 국제보건기구(WHO)에서 제정하며 1629년을 시작으로 계속해서 내용이 보완되고 개정되고 있으며 현재는 ICD-10[1)]를 사용한다. 1992년부터 1998년까지의 자료는 ICD-9를 사망분류 기준으로, 1999년부터 2002년까지의 자료는 ICD-10을 사망분류 기준으로 사용하였다. 사망원인의 하위항목을 살펴보면 직접적인 사인만을 고려한 분류와 간접적인 사인도 함께 고려한 분류를 볼 수 있는데 본 논문에서는 직접적인 사인(항목 4a)만을 고려한다.

주요 질환 선정을 위해 통계청에서 발표한 2004년도 우리나라 사망원인 순위를 참고로 암, 당뇨병, 고혈압, 허혈성심장병, 뇌혈관질환, 하기도질환, 간질환, 교통사고, 자살 및 비만을 10대 사망원인으로 정하였다. 특히 암(악성 신생물)의 경우 종류가 광범위하므로 가장 일반적인 14개의 암(구강 및 인두암, 식도암, 위암, 대장암, 간암, 췌장암, 후두암, 폐암, 유방암, 자궁암, 전립선암, 방광암, 뇌암, 백혈병)을 선정하였다. 또한 사망원인 특성에 맞는 사망자를 대상으로 하여야 하므로 자살은 연령이 5세 이상인 사망자를 대상으로 하였고 유방암, 자궁암 및 전립선암은 각 성별에 맞는 사망자만을 대상으로 분석하였다.

<표 2>는 10대 사망원인과 암에 대한 각 분류기준별 ICD Code에 대해 정리한 표이다. 고혈압, 허혈성 심장병과 같이 질환에 대한 Code가 세분화 되어

1) .ICD-10으로 개정은 단순한 명칭변경 외에 ICD-9와 비교하면 두 가지 측면에서 주요한 변화가 있다.
첫 번째는 종전의 기능에 부가하여 보다 더 상세하게 다른 보건측면을 보완한 다른 의약분류와 연계된 핵심분류이다. 두 번째는 부호체계를 문자와 숫자의 복합구성이다

있는 항목이 존재하여 그와 같은 경우는 세분화된 영역을 그룹화 시켰다.

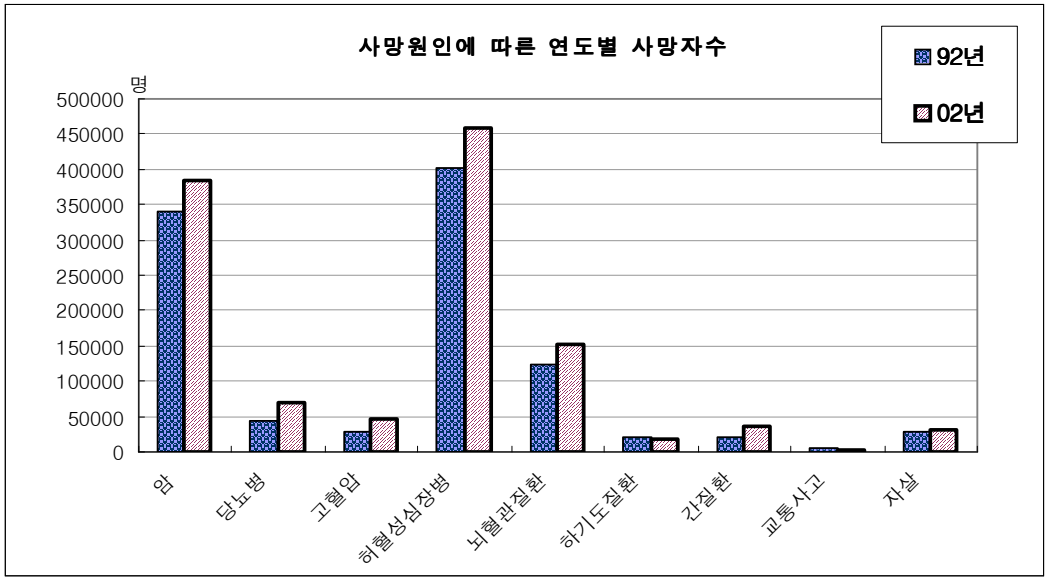
<표 2> 10대 사망원인 및 암에 대한 ICD-Code

| | 당뇨병 | 고혈압 | 허혈성 심장병 | 뇌혈관질환 | 하기도 질환 | 간질환 | 교통사고 | 자살 | 비만 |
|---|---|---|---|---|---|---|---|---|---|
| ICD-9 | 122 | 169-172 | 174-177 | 190-196 | 218-221 | 241 | 301 | 338-344 | 127 |
| ICD-10 | 159 | 206-209 | 211-215 | 235-239 | 264-268 | 298-302 | 386 | 425-431 | 169 |

| | 구강및 인두암 | 식도암 | 위암 | 대장암 | 간암 | 췌장암 | 후두암 | 폐암 | 유방암 | 자궁암 | 전립선암 | 방광암 | 뇌암 | 백혈병 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ICD-9 | 049-052 | 55 | 55 | 058-062 | 065-067 | 69 | 72 | 73 | 80 | 84 | 87 | 91 | 95 | 105-108 |
| ICD-10 | 072-075 | 77 | 78 | 81 | 085-086 | 88 | 92 | 93 | 104 | 107 | 113 | 118 | 122 | 134-137 |

## 2.2 빈도분석

<그림 1> 10대 주요사망원에 대한 92년과 02년도 사망자수



식습관 변화 및 의술의 발달 등과 같은 시간의 변화에 따른 사망원인의 특성을 파악하기 위해 1992년도와 2002년도 자료에 대해서 성별, 연령, 교육수준, 결혼상태에 따른 원인별 사망자수의 변화를 빈도분석을 통해 살펴보고자 한다. 1992년도와 2002년도에 대한 10대 주요사망원인의 전체적인 분포를 알아보기 위하여 <그림 1>에 연도별 사망자수를 그래프로 그려 보았다. 암과 허혈성 심장병으로 인한 사망이 92년과 02년도 모두 가장 높은 빈도를 보였으며 전체적으로 사망원인에 따라 동일한 패턴을 보이고 있음을 알 수 있다. 이 그래프를 기초로 하여 성별, 연령, 결혼상태 및 교육수준에 따른 10대 주요 사망원인의 빈도 및 백분위에 대하여 알아보자.

<표 3> 성별에 따른 1992년과 2002년 사망자 빈도

| | 성별 | 암 | 당뇨병 | 고혈압 | 허혈성 심장병 | 뇌혈관 질환 | 만성 하기도 | 만성 간질환 | 교통 사고 | 자살 | 비만 | 총합 | $X^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 92년 | 남 | 187722 | 18966 | 12104 | 211270 | 49005 | 11078 | 14017 | 3210 | 21761 | 184 | 529317 | 21980.12* |
| | | 55.20 | 43.72 | 42.13 | 52.43 | 39.62 | 52.45 | 64.81 | 70.04 | 80.12 | 37.94 | 52.21 | |
| | 여 | 152330 | 24410 | 16624 | 191718 | 74667 | 10044 | 7612 | 1373 | 5399 | 301 | 484478 | |
| | | 44.80 | 56.28 | 57.87 | 47.57 | 60.38 | 47.55 | 35.19 | 29.96 | 19.88 | 62.06 | 47.79 | |
| 02년 | 남 | 207344 | 32295 | 18958 | 236513 | 58488 | 9276 | 21936 | 2156 | 23933 | 1613 | 612512 | 378985* |
| | | 53.89 | 47.04 | 41.55 | 51.54 | 38.70 | 48.40 | 61.22 | 67.54 | 80.14 | 46.69 | 51.02 | |
| | 여 | 177438 | 36365 | 26669 | 222388 | 92655 | 9890 | 13893 | 1036 | 5931 | 1842 | 588107 | |
| | | 46.11 | 52.96 | 58.45 | 48.46 | 61.30 | 51.60 | 38.78 | 32.46 | 19.86 | 53.31 | 48.98 | |

* $p<0.001$

<표 3>은 성별에 따른 10대 사망원인에 대한 빈도표이다. 각 연도별 성별과 사망원인의 독립성 검정결과 피어슨 카이제곱통계량 $X^2$값이 각각 21980.12, 378985로 유의하게 나타났으며 이는 성별에 따라 각 질환별로 유의한 차이를 보이고 있음을 의미한다. 1992년도를 살펴보면 만성간질환(64.81%), 교통사고(70.04%), 자살(80.12%)은 남자가 높게 나타났으며 뇌혈관질환(60.38%), 비만(62.06%)은 여자가 높게 나타났다. 2002년은 1992년과 전체적으로 비슷한 경향을 보였으나 1992년도에는 남녀의 차이가 많이 나던

비만이 2002년도에는 비슷한 비율로 변하였다. 이는 남자의 경우 육류과식, 과도한 음주, 흡연 등의 건강관리에 관한 부분이 여자보다 불규칙한 생활습관을 갖기 때문에 만성간질환의 사망빈도가 높고 사회적인 활동으로 인하여 교통사고 및 자살의 빈도가 높다고 여겨진다. 반면에 여성의 경우에 뇌혈관 질환 및 비만이 더 높은 이유로 남녀 모두에 적용되는 위험인자 뿐만이 아니라 피임약, 임신, 폐경기 등 여성에게만 적용되는 위험인자들이 존재하기 때문에 남자보다 더 높은 사망을 보인다고 볼 수 있다.

<표 4>은 연령에 따른 10대 사망원인에 대한 빈도표이며 각 연도에 대한 독립성 검정결과 P-값<0.001로 이는 연령계층에 따라 사망원인이 상이함을 나타내고 있다. 연령은 순서형 변수이므로 순서자료에 대한 독립성 검정을 실시한 결과 코크란-맨틀-핸첼통계량($M^2$)값이 각각 2466, 3409로 유의하게 나와 연령이 증가함에 따라 질환별로 사망률에 유의한 차이를 보이고 있다. 1992년과 2002년 모두 교통사고, 자살, 비만을 제외한 나머지 원인에 대해서 연령이 높아질수록 사망자수가 늘어가는 패턴을 보이고 있으며 교통사고와 자살은 40대 이하에서 많이 나타나고 있음을 볼 수 있다. 비만의 경우 1992년도에는 연령이 높아질수록 사망자수가 늘었지만 2002년도에는 오히려 40대 및 50대에 사망자수가 높은 것을 볼 수 있다. 이는 현대에 변화된 식생활이 원인이라 할 수 있는 간편히 먹을 수 있는 인스턴트 식사와 잦은 과식 및 과음이 원인으로 여겨진다.

<표 4> 연령에 따른 1992년과 2002년 사망자 빈도

| | 연령 | 암 | 당뇨병 | 고혈압 | 허혈성 심장병 | 뇌혈관 질환 | 만성 하기도 | 만성 간질환 | 교통 사고 | 자살 | 비만 | 총합 | $X^2$ ($M^2$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 92년 | 10대 이하 | 934 | 52 | 7 | 31 | 112 | 143 | 10 | 689 | 2001 | 3 | 3982 | 365076* (2466)* |
| | | 0.27 | 0.12 | 0.02 | 0.01 | 0.09 | 0.68 | 0.05 | 15.03 | 7.37 | 0.62 | 0.39 | |
| | 20대 | 1182 | 283 | 67 | 322 | 330 | 146 | 155 | 686 | 5108 | 1 | 8280 | |
| | | 0.35 | 0.65 | 0.23 | 0.08 | 0.27 | 0.69 | 0.72 | 14.97 | 18.81 | 0.21 | 0.82 | |
| | 30대 | 4991 | 957 | 384 | 2505 | 1316 | 279 | 1761 | 837 | 5814 | 8 | 18852 | |
| | | 1.47 | 2.21 | 1.34 | 0.62 | 1.06 | 1.32 | 8.14 | 18.26 | 21.41 | 1.65 | 1.86 | |
| | 40대 | 16669 | 1998 | 1145 | 11092 | 3152 | 523 | 3688 | 605 | 4523 | 14 | 43409 | |
| | | 4.90 | 4.61 | 3.99 | 2.75 | 2.55 | 2.48 | 17.05 | 13.20 | 16.65 | 2.89 | 4.28 | |
| | 50대 | 39302 | 4012 | 2325 | 25883 | 5448 | 1448 | 4199 | 423 | 2943 | 21 | 86004 | |
| | | 11.56 | 9.25 | 8.09 | 6.42 | 4.41 | 6.86 | 19.41 | 9.23 | 10.84 | 4.33 | 8.48 | |
| | 60대 | 89418 | 9190 | 4611 | 64753 | 13497 | 4768 | 5643 | 435 | 2747 | 43 | 195105 | |
| | | 26.30 | 21.19 | 16.05 | 16.07 | 10.91 | 22.57 | 26.09 | 9.49 | 10.11 | 8.87 | 19.25 | |
| | 70대 | 110356 | 13781 | 7587 | 117362 | 32840 | 7998 | 4302 | 507 | 2520 | 98 | 297351 | |
| | | 32.45 | 31.77 | 26.41 | 29.12 | 26.55 | 37.87 | 19.89 | 11.06 | 9.28 | 20.21 | 29.33 | |
| | 80대 이상 | 77200 | 13103 | 12602 | 181040 | 66977 | 5817 | 1871 | 401 | 1504 | 297 | 360812 | |
| | | 22.70 | 30.21 | 43.87 | 44.92 | 54.16 | 27.54 | 8.65 | 8.75 | 5.54 | 61.24 | 35.59 | |
| 02년 | 10대 이하 | 944 | 84 | 12 | 47 | 123 | 194 | 49 | 384 | 1698 | 18 | 3553 | 24626.29* (3409)* |
| | | 0.25 | 0.12 | 0.03 | 0.01 | 0.08 | 1.01 | 0.14 | 12.03 | 5.69 | 0.52 | 0.30 | |
| | 20대 | 1061 | 321 | 165 | 330 | 276 | 183 | 190 | 408 | 4695 | 142 | 7771 | |
| | | 0.28 | 0.47 | 0.36 | 0.07 | 0.18 | 0.95 | 0.53 | 12.78 | 15.72 | 4.11 | 0.65 | |
| | 30대 | 4759 | 1119 | 774 | 2779 | 1118 | 363 | 1495 | 458 | 5440 | 399 | 18704 | |
| | | 1.24 | 1.63 | 1.70 | 0.61 | 0.74 | 1.89 | 4.17 | 14.35 | 18.22 | 11.55 | 1.56 | |
| | 40대 | 20268 | 3265 | 2601 | 14384 | 3916 | 752 | 6818 | 583 | 6783 | 797 | 60167 | |
| | | 5.27 | 4.76 | 5.70 | 3.13 | 2.59 | 3.92 | 19.03 | 18.26 | 22.71 | 23.07 | 5.01 | |
| | 50대 | 49682 | 7404 | 4335 | 34248 | 7213 | 1459 | 8419 | 418 | 4717 | 940 | 118835 | |
| | | 12.91 | 10.78 | 9.50 | 7.46 | 4.77 | 7.61 | 23.50 | 13.10 | 15.79 | 27.21 | 9.90 | |
| | 60대 | 81333 | 12005 | 5263 | 56692 | 12674 | 3050 | 6859 | 315 | 2462 | 627 | 181280 | |
| | | 21.14 | 17.48 | 11.53 | 12.35 | 8.39 | 15.91 | 19.14 | 9.87 | 8.24 | 18.15 | 15.10 | |
| | 70대 | 120430 | 19946 | 9178 | 111876 | 34692 | 6415 | 6895 | 343 | 2327 | 368 | 312470 | |
| | | 31.30 | 29.05 | 20.12 | 24.38 | 22.95 | 33.47 | 19.24 | 10.75 | 7.79 | 10.65 | 26.03 | |
| | 80대 이상 | 106305 | 24516 | 23299 | 238545 | 91131 | 6750 | 5104 | 283 | 1742 | 164 | 497839 | |
| | | 27.63 | 35.71 | 51.06 | 51.98 | 60.29 | 35.22 | 14.25 | 8.87 | 5.83 | 4.75 | 41.47 | |

* $p<0.001$

<표 5> 결혼상태에 따른 92년과 02년 사망자 빈도

| | 결혼상태 | 암 | 당뇨병 | 고혈압 | 허혈성심장병 | 뇌혈관질환 | 만성하기도 | 만성간질환 | 교통사고 | 자살 | 비만 | 총합 | $X^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 92년 | 미혼 | 20775 | 3420 | 2388 | 25465 | 8292 | 1372 | 2799 | 1932 | 8564 | 51 | 75058 | 74836.29* |
| | | 6.11 | 7.88 | 8.31 | 6.32 | 6.70 | 6.50 | 12.94 | 42.16 | 31.53 | 10.52 | 7.40 | |
| | 기혼 | 193534 | 19608 | 10517 | 181217 | 47318 | 10081 | 9999 | 1427 | 11211 | 148 | 485060 | |
| | | 56.91 | 45.20 | 36.61 | 44.97 | 38.26 | 47.73 | 46.23 | 31.14 | 41.28 | 30.52 | 47.85 | |
| | 미망인 | 93926 | 16581 | 13150 | 167434 | 59620 | 7411 | 4100 | 539 | 2436 | 253 | 365450 | |
| | | 27.62 | 38.23 | 45.77 | 41.55 | 48.21 | 35.09 | 18.96 | 11.76 | 8.97 | 52.16 | 36.05 | |
| | 이혼 | 31817 | 3767 | 2673 | 28872 | 8442 | 2258 | 4731 | 685 | 4949 | 33 | 88227 | |
| | | 9.36 | 8.68 | 9.30 | 7.16 | 6.83 | 10.69 | 21.87 | 14.95 | 18.22 | 6.80 | 8.70 | |
| 02년 | 미혼 | 25960 | 6362 | 4771 | 34239 | 9976 | 1631 | 5003 | 1307 | 9562 | 971 | 99782 | 12175.58* |
| | | 6.75 | 9.27 | 10.46 | 7.46 | 6.60 | 8.51 | 13.96 | 40.95 | 32.02 | 28.10 | 8.31 | |
| | 기혼 | 202020 | 28088 | 14254 | 177894 | 53676 | 7871 | 15246 | 987 | 11524 | 1403 | 512963 | |
| | | 52.50 | 40.91 | 31.24 | 38.77 | 35.51 | 41.07 | 42.55 | 30.92 | 38.59 | 40.61 | 42.72 | |
| | 미망인 | 108214 | 25912 | 21052 | 198143 | 74263 | 7019 | 6963 | 328 | 2289 | 452 | 444635 | |
| | | 28.12 | 37.74 | 46.14 | 43.18 | 49.13 | 36.62 | 19.43 | 10.28 | 7.66 | 13.08 | 37.03 | |
| | 이혼 | 48588 | 8298 | 5550 | 48625 | 13228 | 2645 | 8617 | 570 | 6489 | 629 | 143239 | |
| | | 12.63 | 12.09 | 12.16 | 10.60 | 8.75 | 13.80 | 24.05 | 17.86 | 21.73 | 18.21 | 11.93 | |

*$p<0.001$

<표 5>은 결혼상태에 따른 10대 사망원인에 대한 빈도표이다. 1992년도와 2002년도에 대하여 독립성 검정결과 피어슨 카이제곱통계량($X^2$)이 각각 74836.29, 12175.58로 통계적으로 유의하며 이는 결혼상태에 따라 사망원인별로 유의한 차이가 존재함을 의미한다. 기혼과 미망인은 1992년도와 2002년도 모두에서 교통사고를 제외한 나머지 원인에 대해 높은 빈도를 차지하고 있으며 1992년 교통사고는 미혼인 사람(42.16%)이 가장 사망빈도가 높았으며 미망인의 경우 고혈압(45.77%)과 뇌혈관질환(48.21%)이 높은 빈도를 보였다. 또한 기혼인 경우에 나머지 원인에 대해 높은 빈도를 보였으며 이혼인 경우는 전체적으로 낮은 빈도를 보였다. 2002년도 교통사고는 미혼인 사람(40.95%)이 가장 사망빈도가 높았으며 미망인의 경우 고혈압(46.14%)과 뇌혈관질환(43.18%), 허혈성심장병(49.13%)이 높은 사망빈도를 보였다. 또한 기혼인 경우에 나머지 원인에 대해 높은 빈도를 보였으며 이혼인 경우는

전체적으로 낮은 빈도를 보였다.

<표 6> 교육수준에 따른 92년과 02년 사망자 빈도

| | 교육기간 | 암 | 당뇨병 | 고혈압 | 허혈성 심장병 | 뇌혈관 질환 | 만성 하기도 | 만성 간질환 | 교통 사고 | 자살 | 비만 | 총합 | $\chi^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 92년 | 8년 미만 | 73051 | 12970 | 8348 | 115960 | 37439 | 5031 | 3913 | 1261 | 2731 | 145 | 260849 | 12928.93* |
| | | 21.48 | 29.90 | 29.06 | 28.78 | 30.27 | 23.82 | 18.09 | 27.51 | 10.06 | 29.90 | 25.73 | |
| | 9-11년 | 46092 | 6463 | 4051 | 53650 | 15676 | 3127 | 3557 | 918 | 4705 | 59 | 138298 | |
| | | 13.55 | 14.90 | 14.10 | 13.31 | 12.68 | 14.80 | 16.45 | 20.03 | 17.32 | 12.16 | 13.64 | |
| | 12년 | 136927 | 15914 | 10468 | 149778 | 43430 | 8388 | 9060 | 1607 | 11342 | 173 | 387087 | |
| | | 40.27 | 36.69 | 36.44 | 37.17 | 35.12 | 39.71 | 41.89 | 35.06 | 41.76 | 35.67 | 38.18 | |
| | 13-15년 | 42842 | 4361 | 3043 | 42701 | 14006 | 2584 | 2876 | 472 | 4635 | 60 | 117580 | |
| | | 12.60 | 10.05 | 10.59 | 10.60 | 11.33 | 12.23 | 13.30 | 10.30 | 17.07 | 12.37 | 11.60 | |
| | 16년 이상 | 41140 | 3668 | 2818 | 40899 | 13121 | 1992 | 2223 | 325 | 3747 | 48 | 109981 | |
| | | 12.10 | 8.46 | 9.81 | 10.15 | 10.61 | 9.43 | 10.28 | 7.09 | 13.80 | 9.90 | 10.85 | |
| 02년 | 8년 미만 | 56650 | 14985 | 9479 | 92646 | 31786 | 3334 | 4869 | 679 | 2015 | 301 | 216744 | 83763.33* |
| | | 14.72 | 21.82 | 20.77 | 20.19 | 21.03 | 17.40 | 13.59 | 21.27 | 6.75 | 8.71 | 18.05 | |
| | 9-11년 | 47556 | 9619 | 5947 | 58325 | 18932 | 2748 | 5178 | 578 | 4356 | 480 | 153719 | |
| | | 12.36 | 14.01 | 13.03 | 12.71 | 12.53 | 14.34 | 14.45 | 18.11 | 14.59 | 13.89 | 12.80 | |
| | 12년 | 165889 | 28259 | 18938 | 193389 | 60875 | 8415 | 15869 | 1251 | 12898 | 1626 | 507409 | |
| | | 43.11 | 41.16 | 41.51 | 42.14 | 40.28 | 43.91 | 44.29 | 39.19 | 43.19 | 47.06 | 42.26 | |
| | 13-15년 | 57252 | 8743 | 6015 | 57930 | 20210 | 2691 | 5531 | 382 | 5739 | 656 | 165149 | |
| | | 14.88 | 12.73 | 13.18 | 12.62 | 13.37 | 14.04 | 15.44 | 11.97 | 19.22 | 18.99 | 13.76 | |
| | 16년 이상 | 57435 | 7054 | 5248 | 56611 | 19340 | 1978 | 4382 | 302 | 4856 | 392 | 157598 | |
| | | 14.93 | 10.27 | 11.50 | 12.34 | 12.80 | 10.32 | 12.23 | 9.46 | 16.26 | 11.35 | 13.13 | |

* $p<0.001$

<표 6>은 교육수준에 따른 10대 사망원인에 대한 빈도표이다. 1992년도 및 2002년도 교육수준에 따른 독립성검정결과 피어슨카이제곱통계량($\chi^2$)값이 각각 12928.93, 83763.33로 통계적으로 유의하며 이는 교육수준에 따라 사망원인간의 유의한 차이를 보임을 의미한다. 여기서 주목할 만한 사실은 1992년도와 2002년도 모두 교육수준이 12년인 경우에 10대 사망원인 모두가 가장 높은 사망빈도를 보이고 있음을 알 수 있으며 교육수준이 8년 미만인 경우에도 높은 사망빈도를 보이고 있음을 알 수 있다. 이는 미국의 경우 사

립학교를 제외하고는 중·고등교육을 의무교육화 하였기에 의무교육을 마친 사람과 그 이상의 기간 동안 교육을 배운 사람들 사이에 사망차이가 존재한다고 보여 진다.

## 2.3 수량화를 이용한 탐색적 방법

2.2절의 빈도분석 결과 성별, 연령, 결혼상태 및 교육수준에 따라 사망원인과의 연관성이 존재한다고 볼 수 있다. 이를 확인하기 위해 2002년도 자료를 이용하여 다변량 자료의 탐색적 방법인 수량화방법Ⅱ를 사용한다.

수량화방법Ⅱ 적용을 위한 반응변수인 10대 사망원인(암, 당뇨병, 고혈압, 허혈성심장병, 뇌혈관질환, 하기도 질환, 간질환, 교통사고, 자살, 비만)과 설명변수인 성별, 연령, 교육수준, 결혼 상태에 대한 범주표는 <표 7>과 같다. 여기서 반응변수에 대한 기준은 비만이며 설명변수들 중 성별은 여자, 연령은 80세 이상, 결혼 상태는 이혼, 교육수준은 16년 이상인 경우를 기준변수로 하여 가변수를 생성하였다.

<표 7> 각 변수들에 대한 범주표

| | 범주1 | 범주2 | 범주3 | 범주4 | 범주5 | 범주6 | 범주7 | 범주8 | 범주9 | 범주10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 사망원인 ($dis\_i$) | 암 | 당뇨병 | 고혈압 | 허혈성 심장병 | 뇌혈관 질환 | 하기도 질환 | 간질환 | 교통 사고 | 자살 | 비만 |
| 성별 ($sex$) | 남자 | 여자 | | | | | | | | |
| 연령 ($age\_i$) | 10대 이하 | 20대 | 30대 | 40대 | 50대 | 60대 | 70대 | 80대 이상 | | |
| 결혼상태 ($mar\_i$) | 미혼 | 기혼 | 미망인 | 이혼 | | | | | | |
| 교육수준 ($edu\_i$) | 8년 미만 | 9-11년 | 12년 | 13-15년 | 16년 이상 | | | | | |

$$dis\_i = \begin{cases} 1\ , & \text{사망원인} = \text{범주 } i \\ 0\ , & \text{그렇지 않으면} \end{cases} \quad (1)$$

$$sex = \begin{cases} 1\ , & \text{성별} = \text{범주 } i \\ 0\ , & \text{그렇지 않으면} \end{cases}$$

$$age\_i = \begin{cases} 1\ , & \text{연령} = \text{범주 } i \\ 0\ , & \text{그렇지 않으면} \end{cases}$$

$$mar\_i = \begin{cases} 1\ , & \text{결혼상태} = \text{범주 } i \\ 0\ , & \text{그렇지 않으면} \end{cases}$$

$$edu\_i = \begin{cases} 1\ , & \text{교육정도} = \text{범주 } i \\ 0\ , & \text{그렇지 않으면} \end{cases}$$

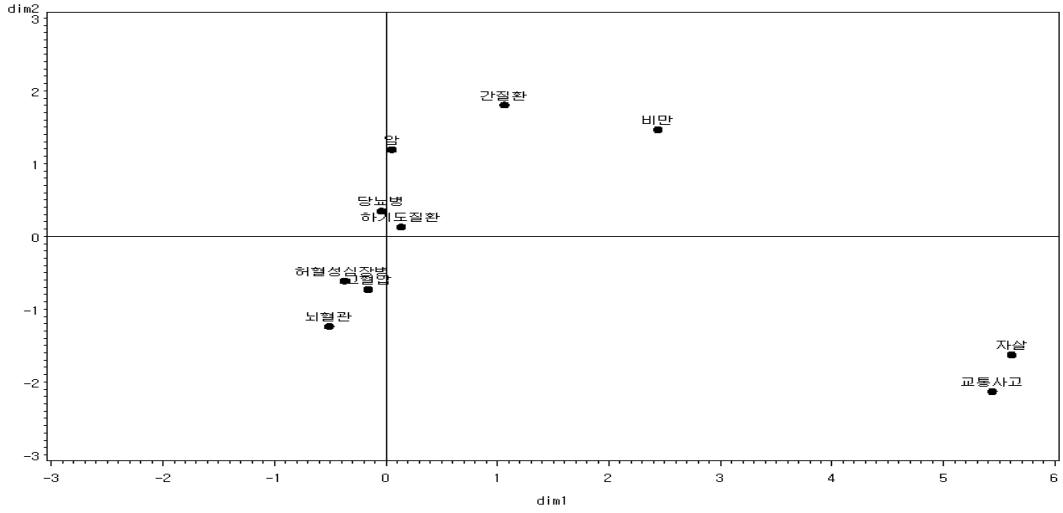
<표 8> 각 변수들에 대한 수량화

| | | 차원1 | 수량화값 | 차원2 | 수량화값 |
|---|---|---|---|---|---|
| 사망원인 | 암 | -2.39 | 0.06 | -0.27 | 1.19 |
| | 당뇨병 | -2.48 | -0.04 | -1.12 | 0.35 |
| | 고혈압 | -2.60 | -0.16 | -2.20 | -0.73 |
| | 허혈성심장병 | -2.81 | -0.37 | -2.08 | -0.61 |
| | 뇌혈관 | -2.95 | -0.51 | -2.70 | -1.24 |
| | 하기도질환 | -2.30 | 0.14 | -1.34 | 0.13 |
| | 간질환 | -1.38 | 1.07 | 0.34 | 1.81 |
| | 교통사고 | 3.00 | 5.44 | -3.60 | -2.13 |
| | 자살 | 3.17 | 5.61 | -3.09 | -1.63 |
| | 비만 | 0.00 | 2.44 | 0.00 | 1.47 |
| 성별 | 남자 | 0.26 | 0.13 | -0.11 | -0.05 |
| | 여자 | 0.00 | -0.13 | 0.00 | 0.06 |
| 교육기간 | 8년미만 | -0.13 | -0.06 | -0.20 | -0.12 |
| | 9-11년 | -0.08 | -0.02 | -0.12 | -0.04 |
| | 12년 | -0.06 | 0.00 | -0.06 | 0.02 |
| | 13-15년 | -0.03 | 0.04 | -0.04 | 0.05 |
| | 16년이상 | 0.00 | 0.06 | 0.00 | 0.08 |
| 연령 | 10대이하 | 7.25 | 6.80 | -0.27 | -1.18 |
| | 20대 | 8.08 | 7.63 | -2.34 | -3.25 |
| | 30대 | 4.18 | 3.74 | 0.11 | -0.80 |
| | 40대 | 1.90 | 1.46 | 1.54 | 0.63 |
| | 50대 | 0.84 | 0.40 | 1.97 | 1.06 |
| | 60대 | 0.43 | -0.02 | 1.96 | 1.05 |
| | 70대 | 0.24 | -0.20 | 1.36 | 0.45 |
| | 80대이상 | 0.00 | -0.45 | 0.00 | -0.91 |
| 결혼상태 | 미혼 | -0.16 | -0.01 | -0.26 | -0.33 |
| | 기혼 | -0.21 | -0.06 | 0.37 | 0.30 |
| | 과부 | -0.14 | 0.02 | -0.17 | -0.25 |
| | 이혼 | 0.00 | 0.15 | 0.00 | -0.07 |
| 정준상관계수 (수정된 상관계수) | | 0.48 (0.48) | | 0.27 (0.27) | |

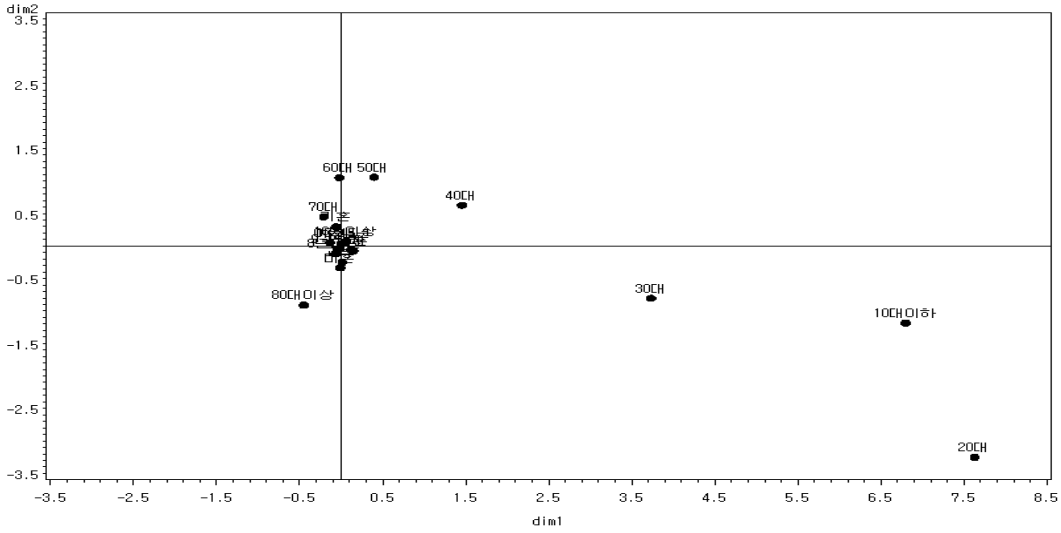
<표 8>은 SAS의 PROC CANCORR를 사용하여 수량화방법Ⅱ를 실행한 결과표이다. 여기서 제3축의 정준상관계수가 상당히 낮게 나타나 제1축과 제2축에 대한 수량화를 고려하였으며, 차원1은 제1축 원정준계수를 의미하고 차원2는 제2축 원정준계수를 의미한다. 차원1에서 사망원인에 대해 뇌혈관질환, 허혈성심장병, 고혈압, 당뇨병, 암, 하기도질환, 간질환은 음수 값을 보이며, 교통사고, 자살은 양수 값을 보이고 있다. 이는 수량화 값이 작아질수록 비상해질환에 의한 사망원인을 나타내며 반대로 수량화 값이 커질수록 상해에 의한 사망원인을 나타낸다. 특히 여자일수록, 교육기간이 짧을수록,

나이가 많을수록, 미혼 혹은 기혼인 경우일수록 비상해로 인하여 사망한다고 볼 수 있으나 여기서 연령을 제외하고는 그 영향력은 매우 작으며, 차원1의 정준상관계수는 0.48로 나타났다. 차원2에서 사망원인에 대한 결과는 교통사고, 자살, 뇌혈관, 고혈압, 심장병의 수량화 값이 음수 값을 보이며 하기도, 당뇨병, 암, 비만, 간질환의 수량화값은 양수이다. 큰 값을 가질수록 만성질환의 성격을 가지지만 정준상관계수가 0.27로 낮게 나타났다.

<그림 2> 사망원인에 대한 수량화 플롯



<그림 3> 설명변수들에 대한 수량화 플롯




<그림 2>와 <그림 3>를 살펴보면 자살과 교통사고와 같은 상해관련 원인으로 나타나는 분류와 비상해 관련된 분류로 나누어지며, 연령이 낮을수록 상해와 관련한 사망원인이 나타나는 경향이 있으며 연령이 높은 80세 이상인 경우에 뇌혈관질환과 같은 비상해로 의한 사망이 나타남을 볼 수 있다. 따라서 1축에 의하여 10대 사망원인의 특징이 상해와 비상해로 구별되어 지며, 제3장에서는 비상해 원인들 중 일반적으로 성인병질환이라 알려진 뇌혈관질환, 심장병, 고혈압 및 당뇨병에 초점을 맞춰서 심도분석을 해보도록 하자.

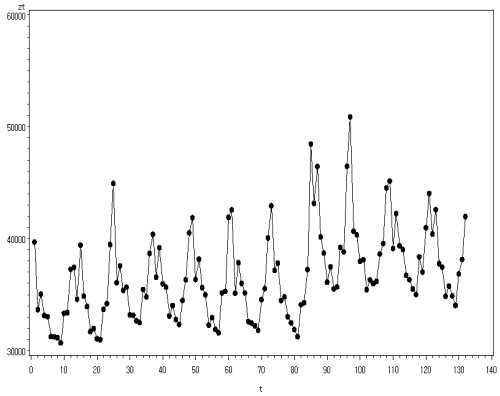
# 제3장 성인병 자료를 이용한 시계열 분석

제2장을 통하여 10대 사망원인들의 일반적인 특징을 알아보았다. 10대 사망원인들 중에서 특징적인 모습을 보인 여러 원인들 중 성인병 관련 질환에 대하여 초점을 맞춰서 심도분석을 하고자 한다. 따라서 대상은 성인병 질환인 허혈성 심장병, 뇌혈관질환, 고혈압, 당뇨병에 대한 1992년 1월부터 2002년 12월까지의 월별 사망자수이다.

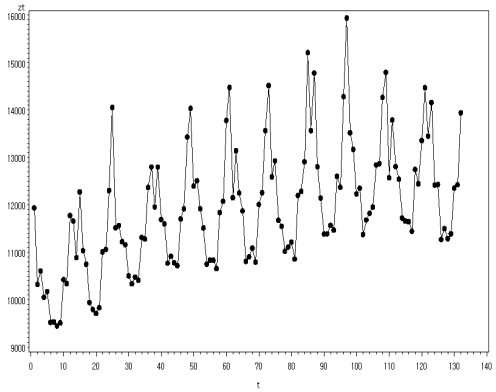
<그림 4>는 1992년 1월부터 2002년 12월까지 월별 허혈성 심장병, 뇌혈관질환, 고혈압 및 당뇨병에 대한 개별 시계열 플롯이다. 질환별 시계열 플롯은 1월에 정점을 이루고 8월에 저점을 이루는 계절주기가 12인 계절성을 가지고 있으며 시간이 경과함에 따라 사망자수가 증가하는 추세를 보이고 있다. 4가지 성인병 시계열 자료에 대해 계절변동을 결정적(deterministic) 또는 확률적(stochastic)으로 간주하여 3가지 분석모형 즉, 선형계절추세모형, 오차항이 ARMA과정을 따르는 회귀모형, 승법계절 ARIMA모형을 적합시켜보고 예측에 유용한 모형을 선택하여 각 질환별 사망자수를 예측하고자 한다.

<그림 4> 성인병 질환들에 대한 개별 시계열 플롯

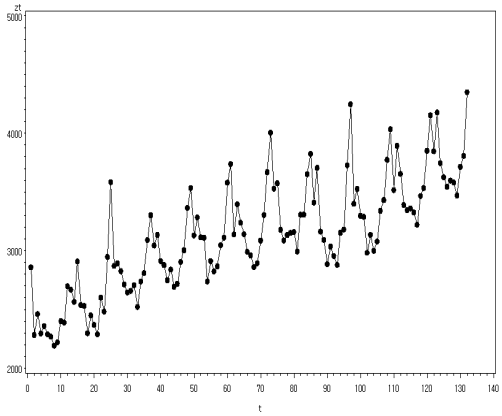
(a) 허혈성 심장병



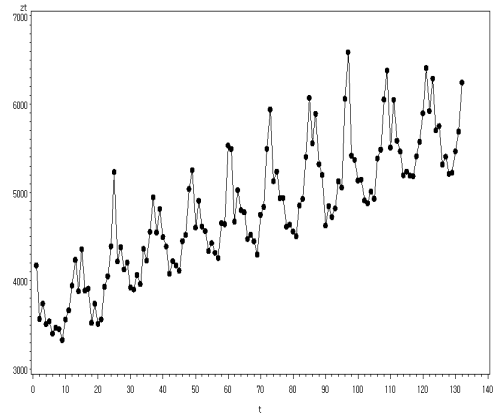

(b) 뇌혈관 질환




(c) 고혈압




(d) 당뇨병



## 3.1 분석모형

### 3.1.1 선형계절추세모형

<그림 4>에서 살펴보았듯이 허혈성 심장병, 뇌혈관질환, 고혈압 및 당뇨병이 선형추세성분과 계절성분을 동시에 갖고 있을 것이라 생각된다(조신섭▪손영숙, 2002). 그러므로 성인병 질환에 대하여 선형계절추세모형을 고려할 수 있으며 이 때의 모형은 식(2)로 표현된다. 모형(2)에서 T_1부터 T_12는 각 월에 해당하는 가변수이며 각 월별 평균을 비교하기 위하여 절편항을 제외시킨 모형을 선택한다. 이 경우 $T_i$에 대한 회귀계수 $\beta_i$는 $i$월의 평균사망자수를 의미한다.

$$z_t = \alpha t + \beta_1 T\_1 + \beta_2 T\_2 + \cdots + \beta_{12} T\_12 + \epsilon_t , \qquad (2)$$

여기서 $\epsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $Cov(\epsilon_t, \epsilon_{t-k}) = 0$,

$t = 1, 2, \cdots, 132$,

$$T\_i = \begin{cases} 1, & i\text{월} \\ 0, & \text{그렇지 않으면} \end{cases} . \qquad (3)$$

### 3.1.2 오차항이 AR과정을 따르는 회귀모형

시계열 자료를 일반회귀모형에 적합시키고자 할 때 회귀모형의 오차항 $\epsilon_t$이 서로 독립인 백색잡음과정을 따르지 않는 경우 오차항이 일반시계열과정을 따르는 회귀모형을 고려해 볼 수 있다.(Yaffee, 2000) 오차항이 AR(p)과정을 따르는 회귀모형은 식(4)로 표현되며 $t$와 가변수에 대한 정의는 모형(2)에서와 같다.

$$z_t = \alpha t + \beta_1 T\_1 + \beta_2 T\_2 + \cdots + \beta_{12} T\_12 + \epsilon_t \ , \qquad (4)$$

여기서 $\epsilon_t = \sum_{j=1}^{p} \phi_j \epsilon_{t-j} + a_t$ ,

$a_t \sim N(0, \sigma_a^2)$ .

### 3.1.3 승법계절 ARIMA $(p, d, q)\times(P, D, Q)_{12}$ 모형

<그림 4>에 주어진 성인병 질환의 계절성분은 다른 추세성분과 독립이라기보다는 확률적으로 볼 수 있다. 또한 시간이 경과함에 따라 성인병 질환 사망에 대한 시계열 $z_t$가 점차 증가하는 추세를 보이는 비정상성을 띠고 있다(이종협▪최기현,2000). 따라서 계절적 특성을 가지고 있는 성인병 질환의 시계열 자료에 대해 계절주기가 12인 승법계절ARIMA $(p, d, q)\times(P, D, Q)_{12}$ 모형이 적합하리라 생각된다. 시계열 $z_t$에서 일반 시계열 모형의 차수를 $p, d, q$라 하고 계절시계열 모형의 차수가 $P, D, Q$라 하면 승법계절 ARIMA 모형은 식(5)와 같다.

$$\Phi(B^{12})\phi(B)(1-B)^d(1-B^{12})^D Z_t = \Theta(B)\Theta(B^{12})a_t \ , \qquad (5)$$

여기서 $\Phi(B^{12}) = 1 - \Phi_1 B^{12} - \Phi_2 B^{24} - \cdots - \Phi_P B^{12P}$ ,

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ ,

$\Theta(B^{12}) = 1 - \Theta_1 B^{12} - \Theta_2 B^{24} - \cdots - \Theta_Q B^{12Q}$ ,

$\Theta(B) = 1 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_q B^q$ ,

$a_t \sim IN(0, \sigma_a^2)$, $E(a_t a_k) = 0$ for $t \neq k$ ,

$E(z_{t-j} a_t) = 0$ for $j > 0$ .

## 3.2 분석결과

### 3.2.1 허혈성심장병

#### 3.2.1.1 선형계절추세모형

선형계절추세모형을 적합시키기 위해 SAS의 PROC REG를 이용하였다. 최소제곱법에 의한 모수추정을 하였고, 오차에 대한 자기상관에 대한 검정으로 Durbin - Watson 검정을 실시한 결과가 <표 9>에 주어져 있다.

<표 9> 모형의 적합결과

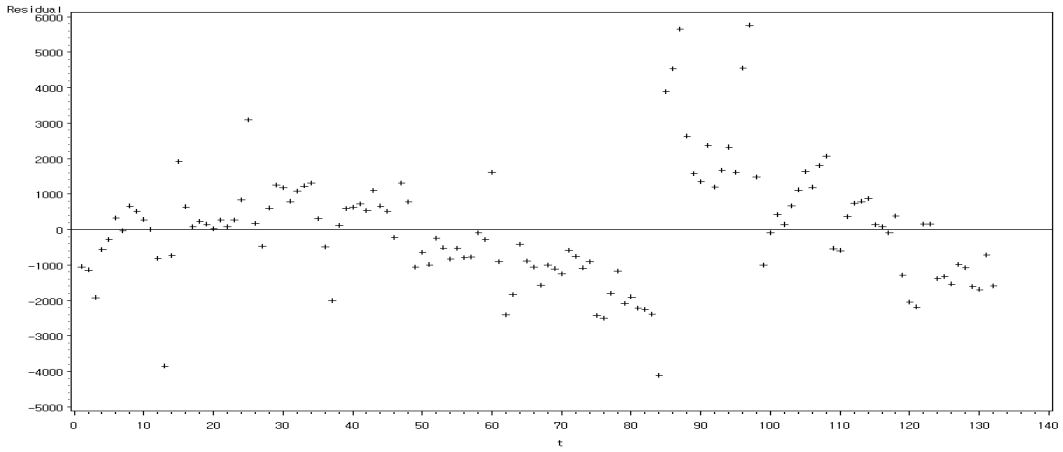
ANOVA

| Source | DF | Sum of Squares | Mean Square | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | 13 | 179067300000 | 13774408429 | 4813.86 | <.0001 |
| Error | 119 | 340507570 | 2861408 | | |
| Uncorrected Total | 132 | 179407800000 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1691.57 | R-Square | 0.9981 |
| Dependent Mean | 36658 | Adj R-Sq | 0.9979 |
| Coeff Var | 4.61446 | 1st Order Autocorrelation | 0.659 |
| Durbin-Watson D | 0.671 | Pr < DW | < 0.001 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 45.60 | 40714 | 34717 | 36835 | 33519 | 33093 | 30663 | 30972 | 30138 | 29776 | 32624 | 32898 | 37554 |
| 표준오차 | 3.88 | 562.27 | 563.91 | 565.58 | 567.26 | 568.97 | 570.70 | 572.45 | 574.23 | 576.02 | 577.83 | 579.67 | 581.52 |

<그림 5> 허혈성 심장병에 대한 선형계절추세모형 $z_t$에 대한 잔차 플롯




<그림 5> 허혈성 심장병에 대한 선형계절추세모형 $z_t$에 대한 잔차 플롯

<표 9>에 주어진 모형의 적합성 검정에 관한 분산분석 결과 유의수준 $\alpha = 0.05$하에서 매우 유의하나 <그림 5>을 살펴보면 선형계절추세모형 적합 후 잔차플롯에서 일정 기간 동안 동일한 부호를 갖는 패턴을 볼 수 있다. 또한 오차항의 자기상관 유무를 검정하기 위한 Durbin-Watson 통계량도 0.671로 P-값이 0.001보다 작아 자기상관이 존재함을 알 수 있다. 실제로 오차의 자기상관계수의 추정값이 0.659로 자기상관이 높은 것으로 나타났다. 따라서 오차가 백색잡음과정을 따른다는 기본가정을 만족시키지 않으며 시간에 따라 관측된 허혈성 심장병 시계열 자료의 경우 오차항이 AR과정을 따르는 회귀모형을 고려해보는 것이 바람직하다.

### 3.2.1.2 오차항이 AR과정을 따르는 회귀모형

모형(2)의 적합 결과 잔차가 양의 상관관계를 갖는 것으로 나타나 오차항의 기본가정을 만족하지 않으므로 $p = 1$인 모형(4)를 적합시키는 것이 바람

직하다. 이를 위해 SAS의 PROC AUTOREG를 이용하였으며 모수추정 방법으로 최우추정법을 사용하였다.

<표 10> 모형의 적합결과

| | | | |
|---|---|---|---|
| SSE | 190870339 | DFE | 118 |
| MSE | 1617545 | Root MSE | 1272 |
| SBC | 2315.8623 | AIC | 2275.503 |
| Regress R-Square | 0.9914 | Total R-Square | 0.9989 |
| Durbin-Watson | 1.8396 | Pr < DW | 0.1892 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 | AR1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 45.22 | 40646 | 34680 | 36818 | 33515 | 33098 | 30672 | 30982 | 30148 | 29782 | 32623 | 32885 | 37522 | -0.6614 |
| 표준오차 | 8.24 | 735.00 | 736.67 | 738.88 | 741.85 | 745.44 | 749.41 | 753.46 | 757.25 | 760.32 | 761.98 | 761.08 | 755.52 | 0.07 |

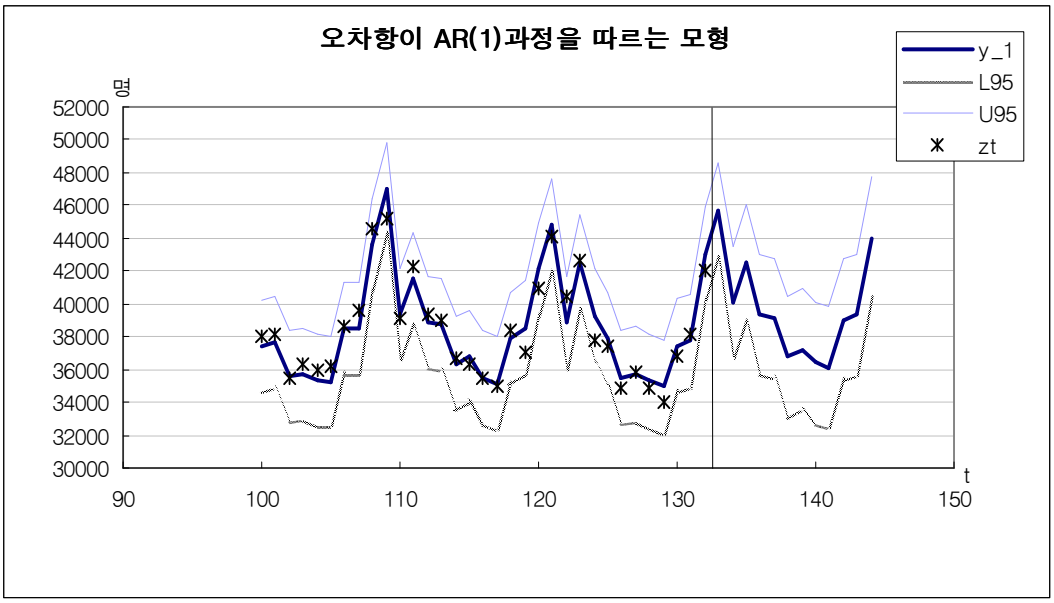
<표 10>는 허혈성 심장병에 대하여 오차항에 AR(1)과정을 적합시킨 모형의 결과 이다. Durbin - Watson 통계량이 1.84이고 P-값이 0.19로 모형의 타당성을 지지하고 있다. 모형(2)에 대한 MSE가 2861408인 반면 모형(4)의 MSE는 1617545로 43.5%정도 감소하였는데 이는 모형의 타당성을 더욱 뒷받침해주는 결과라고 할 수 있다. 따라서 오차항이 AR(1)과정을 따르는 모형의 추정식은

$$\hat{z_t} = 45.22t + 40646\,T\_1 + \cdots + 37522\,T\_12 + \frac{a_t}{1 + 0.66B} \qquad (6)$$

이며 잔차분석 결과 모형의 타당성이 입증되었다. 추정 회귀모형(6)을 기초로 2003년 1월부터 12월까지의 허혈성 심장병으로 인한 사망자수를 예측한 결과가 <그림 6>에 주어져있다. 가로축은 시간($t$)을 나타내며 세로축은 허혈성 심장병으로 인한 사망자수를 나타낸다. $t$가 133에서 144까지가 2003년 한해의 월별 사망자수에 대한 예측값(y_1)을 나타내며 그때의 상한(U95)과 하한(L95)에 대해서도 함께 그렸으며 이때의 자세한 예측 빈도는 부록 A를

참고하길 바란다.

<그림 6> 오차항이 AR(1) 과정을 따르는 모형에 대한 예측 플롯




### 3.2.1.3 승법계절 ARIMA $(p, d, q)\times(P, D, Q)_{12}$모형

계절변동을 확률적으로 간주하여 Box-Jenkins의 승법계절ARIMA $(p,d,q)\times(P,D,Q)_{12}$을 적합시키기 위해 SAS의 PROC ARIMA를 수행하여 모형 식별 통계량을 통해 모형의 차수를 선택한 결과 <표 11>과 같다.

<표 11> 모형 $(1,0,0)\times(2,1,0)_{12}$의 결과

| | MU | $\widehat{\phi_1}$ | $\widehat{\Phi_1}$ | $\widehat{\Phi_2}$ |
|---|---|---|---|---|
| 추정계수 | 471.1086 | 0.6962 | -0.6406 | -0.4298 |
| 표준오차 | 227.3396 | 0.0675 | 0.0871 | 0.0879 |

| | |
|---|---|
| Constant Estimate | 296.3195 |
| Variance Estimate | 2342170 |
| Std Error Estimate | 1530.415 |
| AIC | 2112.704 |
| SBC | 2123.853 |
| Number of Residuals | 120 |

Autocorrelation Check of Residuals

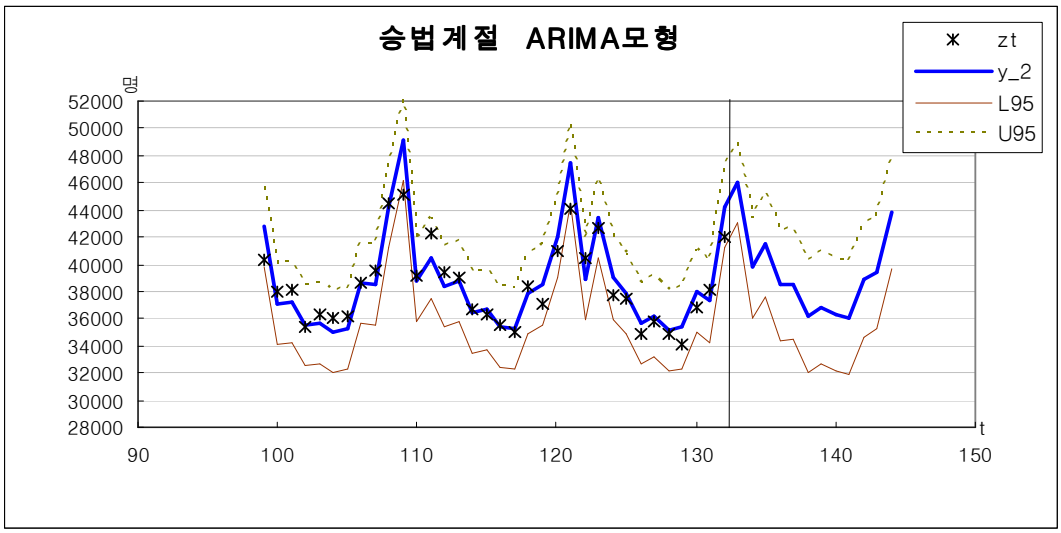
| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 3.52 | 3 | 0.3184 | 0.029 | -0.070 | -0.086 | 0.089 | 0.065 | 0.051 |
| 12 | 14.19 | 9 | 0.1157 | 0.006 | 0.033 | 0.100 | 0.114 | 0.169 | -0.164 |
| 18 | 17.48 | 15 | 0.2909 | -0.143 | 0.042 | 0.038 | -0.020 | 0.009 | -0.001 |
| 24 | 23.44 | 21 | 0.3209 | 0.005 | -0.004 | -0.033 | 0.002 | -0.017 | -0.194 |
| 30 | 26.00 | 27 | 0.5189 | 0.026 | 0.004 | 0.010 | -0.083 | -0.065 | -0.064 |
| 36 | 43.90 | 33 | 0.0972 | -0.024 | -0.094 | 0.052 | -0.165 | -0.062 | -0.247 |

<그림 4a>를 살펴보면 허혈성 심장병이 비정상성을 가지고 있었으며 정상성을 만족하기 위하여 계절차분을 시행하였다. <표 11>은 허혈성 심장병 시계열 $z_t$에 대하여 ARIMA $(1,0,0)\times(2,1,0)_{12}$모형의 적합결과이다. 모수 추정치들은 모두 유의하게 나타났으며, 모형의 적합성을 알아보기 위한 포트맨토우 검정 결과 P-값이0.3184, 0.1157등의 큰 값으로 나타났고 이는 오차가 백색잡음과정임을 지지함으로써 이 모형의 적합성을 잘 나타내주고 있다. 잔차계열에 대한 SACF와 SPACF 역시 오차항이 백색잡음과정임을 보여주고 있다. 따라서 허혈성 심장병 시계열에 대하여 승법계절 ARIMA모형은

$$(1-0.6962B)(1+0.6406B^{12}+0.4297B^{24})(1-B^{12})\widehat{z_t}=296.3195+a_t \qquad (7)$$

이며 이 결과를 가지고 2003년 1월부터 12월까지에 대한 허혈성 심장병으로

인한 사망자수를 예측한다. <그림 7>은 모형(7)을 이용하여 2003년 허혈성 심장병에 대한 사망자수를 예측하여 그래프로 그려 본 결과이다. 가로축은 시간($t$)을 나타내며 세로축은 허혈성 심장병으로 인한 사망자수를 나타낸다. $t$가 133에서 144까지가 2003년 한해의 월별 사망자수에 대한 예측값(y_1)을 나타내며 그때의 상한(U95)와 하한(L95)에 대해서도 함께 그렸으며 이때의 자세한 예측 빈도는 부록 A를 참고하길 바란다.




### 3.2.2 뇌혈관질환

#### 3.2.2.1 선형계절추세모형

뇌혈관 질환의 선형계절추세모형을 적합시키기 위해 SAS의 PROC REG를 이용하였다. 최소제곱법을 이용하여 모수추정을 하였고, 오차에 대한 자기상관관계에 대한 검정으로 Durbin-Watson 검정을 실시한 결과가 <표 12>에 주

어져있다.

<표 12> 모형의 적합결과

ANOVA

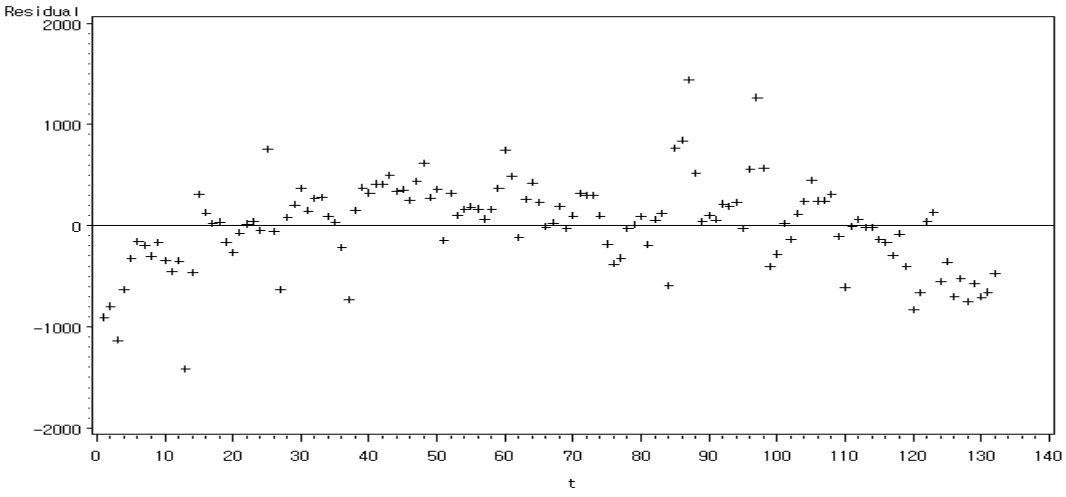
| Source | DF | Sum of Squares | Mean Square | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | 13 | 18996108705 | 1461239131 | 7090.78 | <.0001 |
| Error | 119 | 24523054 | 206076 | | |
| Uncorrected Total | 132 | 19020631759 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 453.9560 | R-Square | 0.9987 |
| Dependent Mean | 11933 | Adj R-Sq | 0.9986 |
| Coeff Var | 3.8041 | 1st Order Autocorrelation | 0.613 |
| Durbin-Watson D | 0.732 | Pr < DW | < 0.001 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 19.10 | 12829 | 11086 | 11685 | 10608 | 10405 | 9566.74 | 9601.73 | 9602.08 | 9506.43 | 10579 | 10586 | 11899 |
| 표준오차 | 1.04 | 150.89 | 151.33 | 151.78 | 152.23 | 152.69 | 153.16 | 153.63 | 154.10 | 154.58 | 155.07 | 155.56 | 156.06 |

<표 12>는 뇌혈관질환에 대한 선형계절추세모형의 분산분석 결과이다. 유의수준 $\alpha = 0.05$하에서 분산분석 결과 모형은 매우 유의하나 <그림 8>을 살펴보면 허혈성 심장병과 같이 일정구간에 동일한 부호를 가짐을 볼 수 있다. 또한 Durbin - Watson 통계량이 0.732로 오차항에 자기상관이 존재한다는 가설을 지지하는 결과이며 P-값이 0.001보다 작아 자기상관이 존재함을 알 수 있다. 실제로 자기상관계수가 0.613으로 높은 양의 상관관계임을 알 수 있다. 따라서 오차가 백색잡음과정을 따른다는 기본가정을 만족시키지 않으며, 시간에 따라 관측된 뇌혈관질환 시계열 자료도 오차항이 AR과정을 따르는 회귀모형을 고려해보는 것이 바람직하다.

<그림 8> 뇌혈관질환에 대한 선형계절추세모형 $z_t$에 대한 오차 플롯




### 3.2.2.2 오차항이 AR과정을 따르는 회귀모형

모형(2)의 적합 결과 잔차가 양의 상관관계를 갖는 것으로 나타나 오차항의 기본가정을 만족하지 않으므로 $p=1$인 모형(4)를 적합시키는 것이 바람직하며 과정은 3.2.1.2절과 같다.

<표 13> 모형의 적합결과

| | | | |
|---|---|---|---|
| SSE | 14880420.5 | DFE | 118 |
| MSE | 126105 | Root MSE | 355.1130 |
| SBC | 1979.0007 | AIC | 1938.6414 |
| Regress R-Square | 0.9944 | Total R-Square | 0.9992 |
| Durbin-Watson | 1.9205 | Pr < DW | 0.3312 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 19.39 | 12786 | 11052 | 11656 | 10583 | 10381 | 9542 | 9576 | 9574 | 9473 | 10538 | 10533 | 11828 | -0.64 |
| 표준오차 | 2.15 | 195.07 | 195.50 | 196.09 | 196.89 | 197.87 | 198.96 | 200.08 | 201.14 | 202.01 | 202.52 | 202.33 | 200.80 | 0.07 |

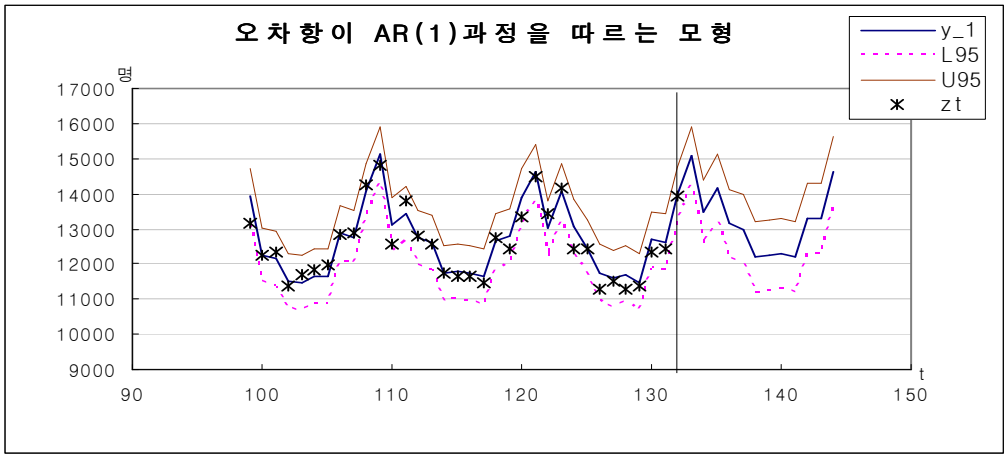
<표 13>는 뇌혈관질환에 대하여 오차항을 AR(1)과정에 적합한 결과이다. Durbin - Watson통계량이 1.9205이고 P-값이 0.33으로 유의수준 $\alpha = 0.05$하에서 유의하지 않아 더 이상 오차항에 자기상관이 존재하지 않는다고 보여진다. 또한 모형(2)에 대한 MSE가 206076이였으나 모형(4)의 MSE는 126105로 약 38.8%정도 감소한 것을 볼 수 있으며 이는 이모형의 타당성을 더욱 뒷받침 해주는 결과라 할 수 있다.

따라서 오차항이 AR(1)과정을 따르는 모형의 추정식은

$$\hat{z_t} = 19.39t + 12786\ T\_1 + \cdots + 11828\ T\_12 + \frac{a_t}{1 + 0.64B} \qquad (8)$$

이며 잔차분석 결과 역시 모형의 타당함을 보여주었다. 추정 회귀모형(8)을 기초로 2003년 1월부터 12월까지의 뇌혈관질환으로 인한 사망자수를 예측한 결과가 <그림 9>이다. 가로축은 시간($t$)을 나타내며 세로축은 뇌혈관질환으로 인한 사망자수를 나타낸다. <그림 7>과 같은 구조이며 이때의 자세한 예측 빈도는 부록 A를 참고하길 바란다.

<그림 9> 오차항이 AR(1) 과정을 따르는 모형에 대한 예측 플롯



### 3.2.2.3 승법계절 ARIMA $(p, d, q)\times(P, D, Q)_{12}$모형

계절변동을 확률적으로 간주하여 Box-Jenkins의 승법계절ARIMA $(p,d,q)\times(P,D,Q)_{12}$을 적합시키기 위해 SAS의 PROC ARIMA를 수행하여 모형 식별 통계량을 통해 모형의 차수를 선택한 결과 <표 14>과 같다.

<표 14> 모형 $(1,0,0)\times(0,1,1)_{12}$의 결과

| | MU | $\widehat{\Theta_1}$ | $\widehat{\phi_1}$ |
|---|---|---|---|
| 추정계수 | 261.6111 | 0.7787 | 0.6623 |
| 표준오차 | 35.8476 | 0.0717 | 0.0720 |

| | |
|---|---|
| Constant Estimate | 88.34147 |
| Variance Estimate | 161042.8 |
| Std Error Estimate | 401.3014 |
| AIC | 1782.238 |
| SBC | 1790.601 |
| Number of Residuals | 120 |

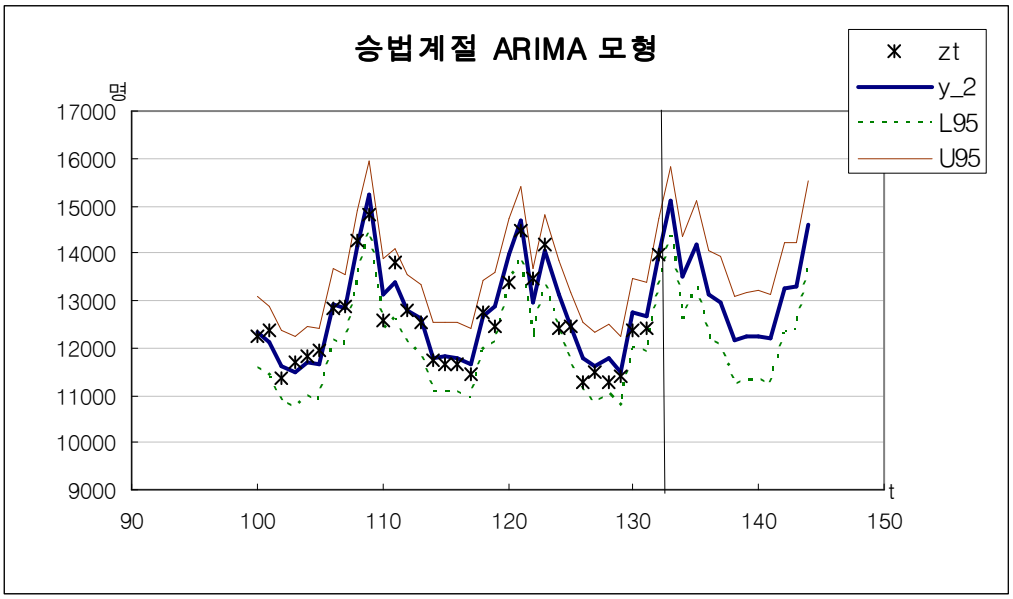
Autocorrelation Check of Residuals

| Lag | To Square | Chi-DF | ChiSq | Pr > | Autocorrelations | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 3.14 | 4 | 0.5346 | 0.022 | -0.111 | -0.007 | 0.075 | 0.067 | -0.044 |
| 12 | 25.26 | 10 | 0.0049 | 0.162 | 0.023 | 0.032 | 0.281 | 0.205 | -0.133 |
| 18 | 31.50 | 16 | 0.0116 | -0.061 | 0.145 | 0.078 | 0.003 | 0.026 | 0.115 |
| 24 | 34.66 | 22 | 0.0421 | 0.011 | 0.016 | 0.002 | 0.089 | -0.077 | 0.082 |
| 30 | 38.69 | 28 | 0.0861 | 0.070 | 0.125 | 0.008 | -0.062 | 0.017 | -0.034 |

<그림 4b>를 살펴보면 뇌혈관질환이 계절에 대하여 비정상성을 보이며 정상성을 만족하기 위하여 계절차분을 시행하였다. <표 14>은 뇌혈관질환 시계열 $z_t$에 대하여 ARIMA $(1,0,0)\times(0,1,1)_{12}$모형의 결과이다. 모형의 모수 추정치들이 모두 유의하였으며 모형의 적합성을 알아보기 위한 포트맨토우 검정결과 P-값이 0.5346의 큰 값으로 나타났고 이는 오차가 백색잡음

과정임을 지지함으로써 뇌혈관질환에 대한 이 모형에 대한 적합성을 잘 나타내주고 있다. 잔차계열에 대한 SACF와 SPACF도 오차항이 백색잡음과정임을 보여주고 있다.

<그림 10> 승법계절 ARIMA $(1,0,0)\times(0,1,1)_{12}$모형에 대한 예측 플롯




따라서 뇌혈관질환 시계열에 대하여 승법계절 ARIMA모형은

$$(1-0.6623B)(1-B^{12})\hat{z_t}=88.3415+(1-0.7787B^{12})a_t \qquad (9)$$

이며 이 결과로 2003년 1월부터 12월까지의 뇌혈관질환으로 인한 사망자수를 예측하려한다. <그림 10>은 2003년 뇌혈관질환에 대한 사망자수를 예측하여 그래프로 그려 본 것이다. 그림에 대한 형식은 <그림 7>과 같으며 이때의 자세한 예측 빈도는 부록 A를 참고하길 바란다.

### 3.2.3 고혈압

#### 3.2.3.1 선형계절추세모형

선형계절추세모형을 적합시키기 위해 3.2.1.1절과 같은 방법을 시행하였으며 최소제곱법에 의한 모수추정 및 Durbin-Watson 검정 결과는 <표 15>와 같다.

<표 15> 모형의 적합결과

ANOVA

| Source | DF | Sum of Squares | Mean Square | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | 13 | 1317381731 | 101337056 | 5326.64 | <.0001 |
| Error | 119 | 2263924 | 19025 | | |
| Uncorrected Total | 132 | 1319645655 | | | |

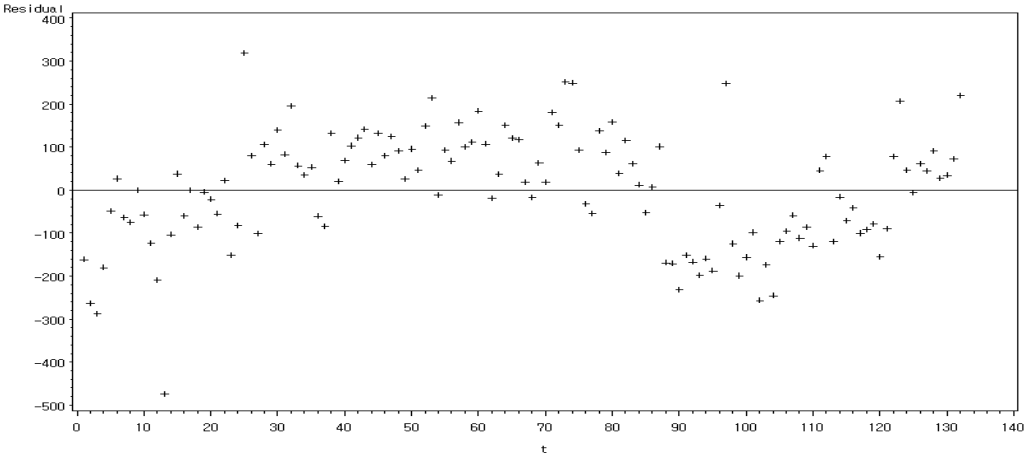
| | | | |
|---|---|---|---|
| Root MSE | 137.92959 | R-Square | 0.9983 |
| Dependent Mean | 3125.7197 | Adj R-Sq | 0.9981 |
| Coeff Var | 4.41273 | 1st Order Autocorrelation | 0.599 |
| Durbin-Watson D | 0.768 | Pr < DW | < 0.001 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 10.19 | 3011.48 | 2527.20 | 2718.65 | 2436.64 | 2358.45 | 2201.54 | 2261.35 | 2187.07 | 2130.33 | 2357.24 | 2401.78 | 2786.22 |
| 표준오차 | 0.32 | 45.85 | 45.98 | 46.12 | 46.25 | 46.39 | 46.53 | 46.68 | 46.82 | 46.97 | 47.12 | 47.27 | 47.42 |

<표 15>에 주어진 모형의 적합성 검정에 관한 분산분석 결과 유의수준 $\alpha = 0.05$하에서 매우 유의하나 <그림 11>을 살펴보면 잔차가 일정한 기간 동안 동일한 부호를 갖는 패턴을 볼 수 있다. 또한 Durbin - Watson통계량도 0.768로 P-값이 0.001보다 작아 자기상관이 존재함을 알 수 있으며 오차의 자기상관계수가 0.599로써 양의 자기상관을 보여준다. 따라서 오차항이 백색잡음과정을 따른다는 기본가정을 만족하지 않으므로 시간에 따라 관측된 고혈압 시계열 자료 역시 오차항이 AR과정을 따르는 회귀모형을 고려해보는 것이

바람직하다.

<그림 11> 고혈압에 대한 선형계절추세모형 $z_t$에 대한 잔차 플롯




### 3.2.3.2 오차항이 AR과정을 따르는 회귀모형

모형(2)의 적합 결과 잔차가 양의 상관관계를 갖는 것으로 나타나 오차항의 기본가정을 만족하지 않으므로 $p=1$인 모형(4)를 적합시키는 것이 바람직하며 이를 위해 SAS의 PROC AUTOREG를 이용하고, 모수추정방법으로 최우추정법을 사용하였다.

<표 16> 모형의 적합결과

| | | | |
|---|---|---|---|
| SSE | 1416963 | DFE | 118 |
| MSE | 12008 | Root MSE | 109.5818 |
| SBC | 1668.565 | AIC | 1628.206 |
| Regress R-Square | 0.993 | Total R-Square | 0.9989 |
| Durbin-Watson | 2.029 | Pr < DW | 0.5631 |

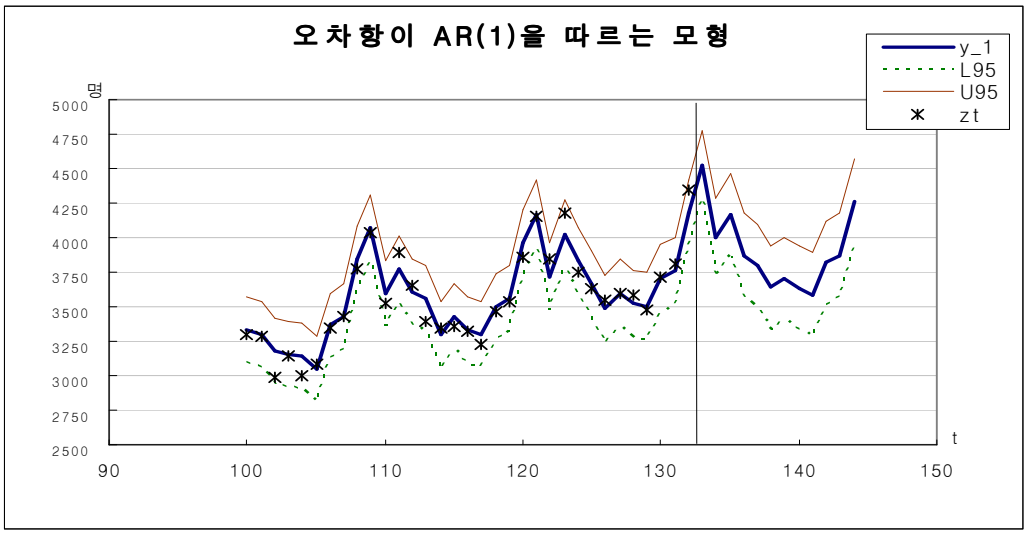
| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 | AR1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 10.42 | 3009 | 2520 | 2709 | 2425 | 2345 | 2187 | 2246 | 2170 | 2112 | 2338 | 2380 | 2761 | -0.62 |
| 표준오차 | 0.64 | 58.46 | 58.58 | 58.76 | 59.01 | 59.31 | 59.65 | 59.99 | 60.32 | 60.60 | 60.76 | 60.72 | 60.26 | 0.07 |

<표 16>은 고혈압에 대한 오차항이 AR(1)과정인 모형의 적합 결과이다. Durbin - Watson통계량이 1.92이고 P-값이 0.56로 모형이 타당함을 지지하고 있다. 또한 모형(2)의 MSE가 19025였으나 모형(4)의 MSE가 12008로 36.9% 감소하였으며 이는 이 모형의 타당성을 더욱 뒷받침 해주는 결과이다. 따라서 오차항이 AR(1)과정을 따르는 모형의 추정식은

$$\hat{z_t} = 10.42t + 3009\ T\_1 + \cdots + 2761\ T\_12 + \frac{a_t}{1 + 0.62B} \qquad (10)$$

이며 잔차분석 결과도 모형의 타당성을 뒷받침하였다. <그림 12>은 추정 회귀모형(10)을 기초로 2003년 1월부터 12월까지 고혈압 사망자수를 예측하여 그래프로 그린 것이다. 가로축은 시간($t$)을 나타내며 세로축은 고혈압으로 인한 사망자수를 나타낸다. 자세한 예측 빈도는 부록 A를 참고하길 바란다.

<그림 12> 오차항이 AR(1)과정을 따르는 모형에 대한 예측 플롯



### 3.2.3.3 승법계절 ARIMA $(p, d, q)\times(P, D, Q)_{12}$모형

계절변동을 확률적으로 간주하여 Box-Jenkins의 승법계절ARIMA $(p,d,q)\times(P,D,Q)_{12}$을 적합시키기 위해 3.2.1.3과 같은 과정을 시행한 결과가 <표 17>과 같다.

<표 17> 모형 $(1,1,1)\times(0,1,1)_{12}$의 결과

| | $\widehat{\theta}_1$ | $\widehat{\Theta}_1$ | $\widehat{\phi}_1$ |
|---|---|---|---|
| 추정계수 | 0.7633 | 0.6885 | 0.4197 |
| 표준오차 | 0.1197 | 0.0750 | 0.1694 |

| | |
|---|---|
| Variance Estimate | 18429.16 |
| Std Error Estimate | 135.7541 |
| AIC | 1509.45 |
| SBC | 1517.787 |
| Number of Residuals | 119 |

Autocorrelation Check of Residuals

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 2.19 | 3 | 0.5336 | 0.045 | -0.113 | 0.017 | -0.030 | 0.040 | 0.003 |
| 12 | 14.51 | 9 | 0.1053 | -0.025 | -0.032 | 0.022 | 0.239 | 0.173 | -0.062 |
| 18 | 18.98 | 15 | 0.2145 | 0.118 | 0.115 | 0.019 | -0.041 | -0.025 | 0.053 |
| 24 | 24.81 | 21 | 0.2555 | -0.071 | -0.057 | 0.037 | 0.037 | 0.159 | -0.054 |
| 30 | 30.30 | 27 | 0.3009 | 0.104 | 0.052 | 0.034 | -0.001 | -0.118 | -0.079 |
| 36 | 34.22 | 33 | 0.4090 | 0.034 | 0.056 | 0.014 | 0.014 | 0.039 | 0.129 |

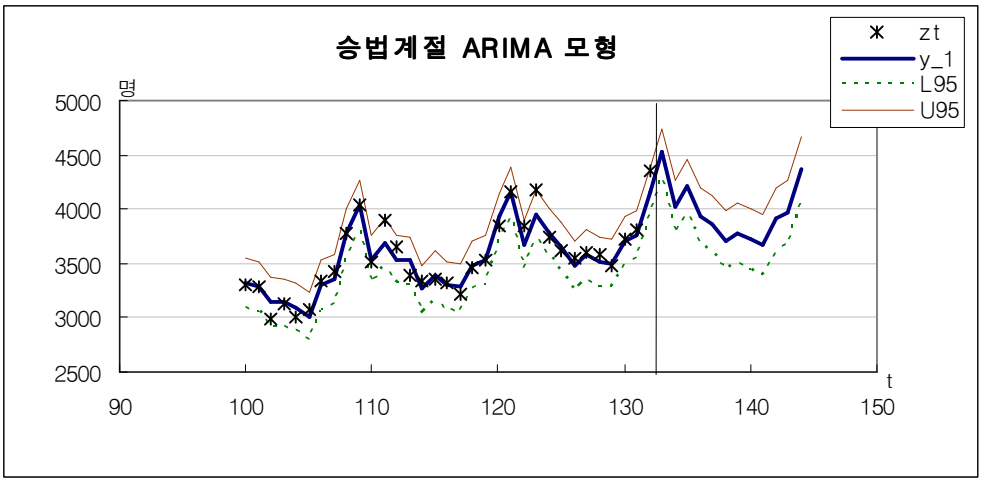
<그림 4c>를 살펴보면 고혈압이 일반 및 계절에 대하여 비정상성을 가지고 있었으며 정상성을 만족하기 위하여 1차 차분 및 계절차분을 하였다. <표 17>은 고혈압 시계열 $z_t$에 대한 ARIMA $(1,1,1)\times(0,1,1)_{12}$모형의 적합결과이다. 모든 모수 추정치들이 유의하였으며 모형의 적합성을 알아보기 위한 포트맨토우 검정결과를 살펴보면 P-값이 0.5336의 큰 값으로 나타났다. 이는 오차가 백색잡음과정임을 지지함으로써 이 모형의 적합성을 잘 나타내고 있

으며 잔차계열에 대한 SACF 및 SPACF 역시 오차항이 백색과정임을 잘 보여주고 있다. 따라서 고혈압 시계열에 대하여 승법계절 ARIMA모형은

$$(1-0.41969B)(1-B)(1-B^{12})\hat{z_t}=(1-0.76326B)(1-0.68845B^{12})a_t \qquad (11)$$

이며 <그림 13>은 이 결과를 이용하여 2003년 1월부터 12월까지의 고혈압에 대한 사망자수를 예측하여 그래프로 그린 것이다. 자세한 예측 빈도는 부록 A를 참고하길 바란다.

<그림 13> 승법계절 ARIMA $(1,1,1)\times(0,1,1)_{12}$모형에 대한 예측 플롯




### 3.2.4 당뇨병

#### 3.2.4.1 선형계절추세모형

선형계절추세모형을 적합시키기 위하여 3.2.1.1절과 같은 방법을 시행하였

으며 최소제곱법에 의한 모수추정 및 Durbon-Watson 검정결과는 <표 18>과 같다.

<표 18> 모형의 적합결과

ANOVA

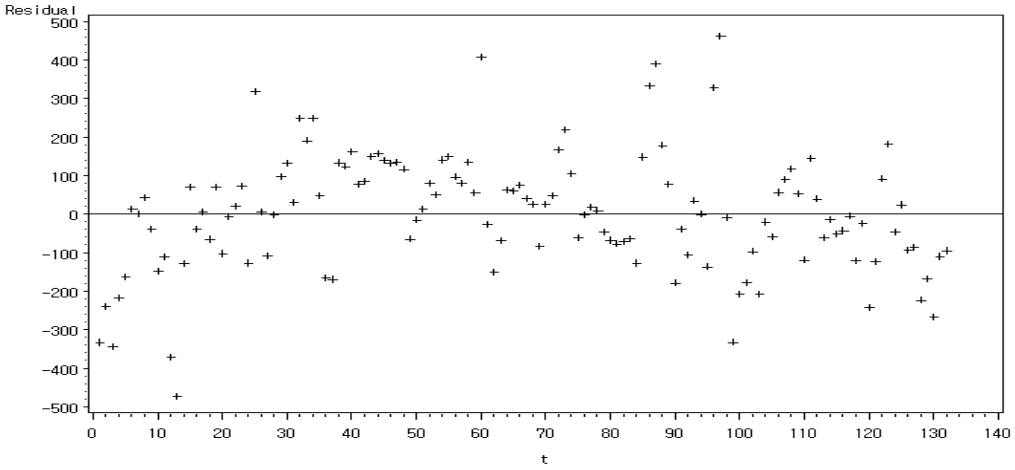
| Source | DF | Sum of Squares | Mean Square | F Value | Pr>F |
|---|---|---|---|---|---|
| Model | 13 | 3094581676 | 238044744 | 8982.68 | <.0001 |
| Error | 119 | 3153550 | 26500 | | |
| Uncorrected Total | 132 | 3097735226 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 162.78949 | R-Square | 0.999 |
| Dependent Mean | 4786.5 | Adj R-Sq | 0.9989 |
| Coeff Var | 3.40101 | 1st Order Autocorrelation | 0.524 |
| Durbin-Watson D | 0.914 | Pr < DW | < 0.001 |

| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 16.87 | 4493.82 | 3776.50 | 4035.36 | 3660.12 | 3621.98 | 3288.11 | 3351.79 | 3278.74 | 3218.51 | 3543.27 | 3592.59 | 4115.90 |
| 표준오차 | 0.37 | 54.11 | 54.27 | 54.43 | 54.59 | 54.76 | 54.92 | 55.09 | 55.26 | 55.43 | 55.61 | 55.78 | 55.96 |

<표 18>에 주어진 모형의 적합성검정에 관한 분산분석 결과 유의수준 $\alpha = 0.05$하에서 매우 유의하나 <그림 14>을 살펴보면 일정한 기간 동안 동일한 부호를 갖는 패턴이 존재함을 볼 수 있다. 또한 오차항의 자기상관 유무를 검정하기 위한 Durbin-Watson통계량도 0.914로 P-값이 0.001보다 작아 자기상관이 존재함을 알 수 있으며, 오차의 자기상관계수의 추정값이 0.524로 이는 어느 정도 양의 자기상관이 존재할 것으로 생각되어진다. 따라서 오차가 백색잡음과정을 따른다는 기본가정을 만족시키지 않으며 시간에 따른 당뇨병 시계열 자료의 경우 오차항이 AR과정을 따르는 회귀모형을 고려해보는 것이 바람직하다.

<그림 14> 당뇨병에 대한 선형계절추세모형 $z_t$에 대한 잔차 플롯




### 3.2.4.2 오차항이 ARMA과정을 따르는 회귀모형

모형(2)의 적합 결과 잔차가 양의 상관을 갖는 것으로 나타나 오차항의 기본가정을 만족하지 않으므로 $p=1$인 모형(4)를 적합시키는 것이 바람직하며 3.2.1.2절과 같은 방법으로 모수추정 및 모형적합시킨 결과가 <표 19>에 주어져 있다.

<표 19> 모형의 적합결과

| | | | |
|---|---|---|---|
| SSE | 2250733.43 | DFE | 118 |
| MSE | 19074 | Root MSE | 138.1087 |
| SBC | 1729.5099 | AIC | 1689.1507 |
| Regress R-Square | 0.9967 | Total R-Square | 0.9993 |
| Durbin-Watson | 1.8913 | Pr < DW | 0.2744 |

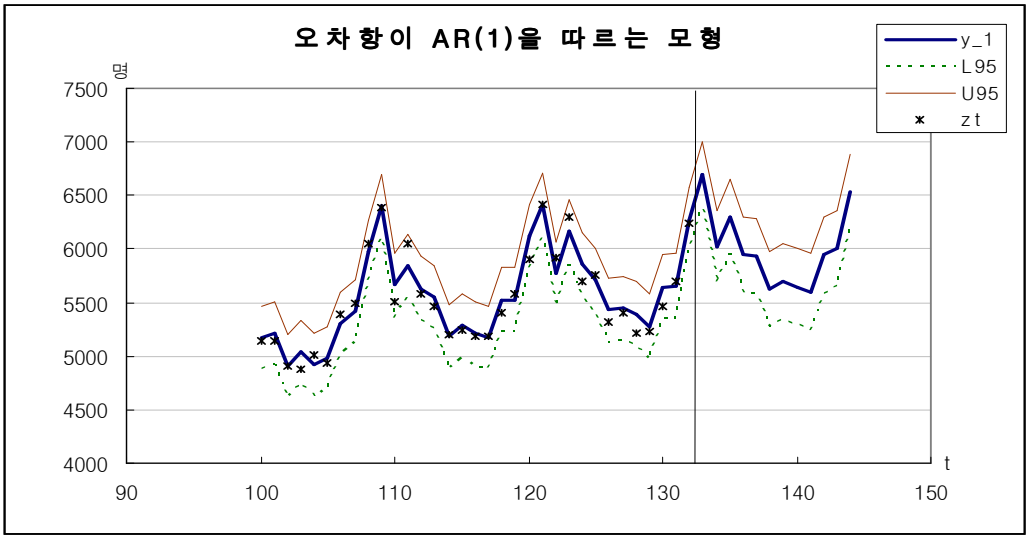
| 변수 | $t$ | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 | T_8 | T_9 | T_10 | T_11 | T_12 | AR1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 추정계수 | 16.97 | 4483 | 3768 | 4027 | 3653 | 3615 | 3281 | 3344 | 3270 | 3209 | 3531 | 3576 | 4092 | -0.54 |
| 표준오차 | 0.67 | 65.25 | 65.38 | 65.57 | 65.86 | 66.20 | 66.58 | 66.97 | 67.35 | 67.69 | 67.92 | 67.93 | 67.42 | 0.08 |

<표 16>은 당뇨병에 대한 오차항이 AR(1)과정인 모형의 적합 결과이다. 또한 Durbin - Watson 통계량이 1.92, P-값이 0.56로 오차항에 더 이상 자기상관이 존재하지 않음을 알 수 있다. 또한 모형(2)의 MSE가 26500이였으나 모형(4)의 MSE가 19074로 약28.0% 감소하였는데 이는 이 모형의 타당성을 더욱 뒷받침 해주는 결과이다. 따라서 오차항이 AR(1)과정을 따르는 모형의 추정식은

$$\hat{z_t} = 16.97t + 4483\ T\_1 + \cdots + 4092\ T\_12 + \frac{a_t}{1 + 0.54\,B} \qquad (12)$$

이며 잔차분석 결과 모형의 타당성이 입증되었다. 추정 회귀모형(12)를 기초로 2003년 1월부터 12월까지의 당뇨병으로 인한 사망자수를 예측한 결과가 <그림 15>이다. 자세한 예측 빈도는 부록 A를 참고하길 바란다.

<그림 15> 오차항이 AR(1) 과정을 따르는 모형에 대한 예측 플롯



### 3.2.4.3 승법계절모형 ARIMA $(p, d, q)\times(P, D, Q)_{12}$

계절변동을 확률적으로 간주하여 Box-Jenkins의 승법계절ARIMA $(p,d,q)\times(P,D,Q)_{12}$을 적합시키기 위해 3.2.1.3과 같은 과정을 시행한 결과가 <표 20>이다.

<표 20> 모형 $(1,1,1)\times(0,1,1)_{12}$의 결과

| | $\widehat{\Theta}_1$ | $\widehat{\Theta}_1$ | $\widehat{\Phi}_1$ |
|---|---|---|---|
| 추정계수 | 0.9217 | 0.8109 | 0.4264 |
| 표준오차 | 0.0580 | 0.1002 | 0.1032 |

| | |
|---|---|
| Variance Estimate | 22119.79 |
| Std Error Estimate | 148.7272 |
| AIC | 1545.721 |
| SBC | 1554.059 |
| Number of Residuals | 119 |

Autocorrelation Check of Residuals

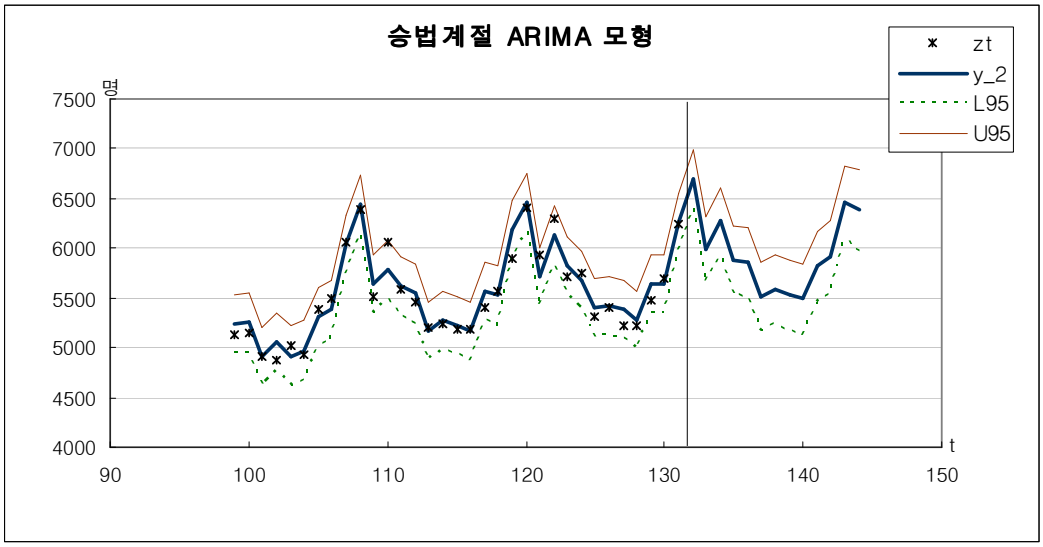
| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 6.15 | 3 | 0.1046 | 0.069 | -0.072 | -0.150 | 0.017 | -0.120 | -0.047 |
| 12 | 15.28 | 9 | 0.0836 | -0.079 | 0.071 | 0.067 | 0.203 | 0.104 | -0.038 |
| 18 | 18.73 | 15 | 0.2260 | 0.003 | 0.039 | 0.133 | -0.071 | -0.025 | -0.006 |
| 24 | 20.51 | 21 | 0.4892 | 0.038 | 0.072 | 0.037 | 0.004 | 0.011 | -0.062 |
| 30 | 26.58 | 27 | 0.4865 | 0.107 | 0.165 | -0.001 | -0.001 | 0.024 | -0.011 |

<그림 4d>를 살펴보면 당뇨병이 비정상성을 가지고 있었으며 정상성을 만족하기 위하여 1차 차분 및 계절차분을 시행하였다. <표 20>은 당뇨병 시계열 $z_t$에 대하여 모형 $(1,1,1)\times(0,1,1)_{12}$의 결과이다. 모든 모수의 추정치들은 유의하였으며, 모형의 적합성을 알아보기 위한 포트맨토우 검정결과 P-값이 0.1046등의 큰 값으로 나타났고, 이는 오차가 백색잡음과정임을 지지함으로써 이모형의 적합성을 잘 나타내주고 있다. 따라서 당뇨병 시계열에 대하여 승법계절 ARIMA모형은

$$(1-0.4264B)(1-B)(1-B^{12})z_t = (1-0.9217B)(1-0.8109B^{12})a_t \qquad (13)$$

이며 모형(13)을 이용하여 2003년 1월부터 12월까지의 당뇨병으로 인한 사망자수를 예측한 결과가 <그림 16>에 주어져 있다. 자세한 예측 빈도는 부록 A를 참고하길 바란다.

<그림 16> 승법계절 ARIMA $(1,1,1)\times(0,1,1)_{12}$모형에 대한 예측 플롯




## 3.3 요약

허혈성 심장병, 뇌혈관질환, 고혈압 및 당뇨병의 시계열 자료에 대해 선형계절추세모형, 오차항이 AR(1)과정을 따르는 회귀모형 및 승법계절 ARIMA모형을 적용시켜 보았다.

그 결과를 종합하면 모든 성인병 질환에 대하여 선형계절추세모형은 잔차들간의 양의 자기상관이 높은 것으로 나타나 오차항의 기본가정을 만족하지 않으므로 이 모형은 적합하지가 않았다.

오차항이 AR(1)과정을 따르는 회귀모형에서는 네 가지의 성인병 질환에 대한 결과가 비슷하게 나타났는데, Durbin - Watson통계량이 2에 가까운 값으로 나타나 더 이상 오차항에 자기상관이 존재하지 않음을 알 수 있었다. 허혈성심장병, 뇌혈관질환, 고혈압 및 당뇨병에 각각 해당되는 모형(6), 모형(8), 모형(10), 모형(12)의 추정계수도 모두 유의하게 나타났으며, 시간이 경과할수록 사망자수는 증가하고 있으며 6월부터 9월까지가 다른 월에 비해 상대적으로 적게 발병하는 것으로 나타났다. 또한 사망발병이 9월이 가장 적고 1월이 가장 높게 나타나는 것으로 보아 겨울철에 특히 성인병질환으로 인한 사망에 대하여 유의해야 할 것으로 보인다.

승법계절 ARIMA모형의 결과를 살펴보면 허혈성 심장병과 뇌혈관질환은 계절에 대하여 비정상성을 갖고 있었으며 허혈성 심장병에 대한 모형(7)에서 $t$시점의 월별 사망자수는 $(t-1)$시점에서의 사망자수뿐만 아니라 $(t-12)$, $(t-13)$시점과 $(t-24)$,$(t-25)$시점에 걸쳐 영향을 받고 있었다. 뇌혈관질환의 경우 모형(9)에서 t시점의 월별 사망자수는 $(t-1)$시점에서의 사망자수와 전년도인 $(t-12)$, $(t-13)$시점에 걸쳐 영향을 받고 있었다. 반면 고혈압과 당뇨병의 경우 각 시계열이 비정상성을 갖고 있었으며 계절주기에 대하여도 비정상성을 갖고 있었다. t시점의 월별 사망자수는 $(t-1)$시점뿐만이 아니라 설명할 수 없는 변동들의 선형결합도 존재하였으므로 그 패턴을 단정지어 설명할 수는 없어 보인다.

3장에서 제시한 세 가지 모형 중 적합하게 나타난 오차항이 AR(1)과정을 따르는 회귀모형과 승법계절 ARIMA모형에 대해 두 모형의 AIC와 SBC를 비교

해보자. <표 21>은 허혈성심장병, 뇌혈관질환, 고혈압 및 당뇨병에 대한 두 모형의 AIC와 SBC를 정리한 표이다. 각 성인병 질환에 대하여 오차항이 AR(1)과정을 따르는 회귀모형의 AIC 및 SBC보다 승법계절 ARIMA모형에서의 AIC 및 SBC가 작은 값을 나타내었다. 이는 동일한 자료에 대한 모형별 AIC 및 SBC이므로 비교 가능하며 승법계절 ARIMA모형에 대하여 AIC와 SBC가 더 작은 값으로 나타났으므로 두 모형 중 승법계절ARIMA모형을 이용하는 것이 더 적절할 것으로 보인다.

<표 21> 두 모형의 AIC와 SBC 비교

| | 오차항이 AR(1)과정을 따르는 회귀모형 | | 승법계절 ARIMA모형 | |
|---|---|---|---|---|
| | AIC | SBC | AIC | SBC |
| 허혈성심장병 | 2275.503 | 2315.862 | 2112.704 | 2123.853 |
| 뇌혈관질환 | 1938.641 | 1979.001 | 1782.238 | 1790.601 |
| 고혈압 | 1628.206 | 1668.565 | 1509.45 | 1517.787 |
| 당뇨병 | 1689.151 | 1729.510 | 1545.721 | 1554.059 |

따라서 허혈성심장병은 승법계절 $(1,0,0)\times(2,1,0)_{12}$모형, 뇌혈관질환은 승법계절 $(1,0,0)\times(0,1,1)_{12}$모형, 고혈압은 승법계절 $(1,1,1)\times(0,1,1)_{12}$모형, 당뇨병은 승법계절 $(1,1,1)\times(0,1,1)_{12}$모형이 예측모형으로 적절하다고 볼 수 있다.

# 제4장 결론

본 논문은 미국의 사망자료를 이용하여 주요 사망원인을 정하여 성별, 연령별, 결혼상태 및 교육수준에 따른 빈도분석과 수량화 방법을 이용한 기초분석을 실시하였다.

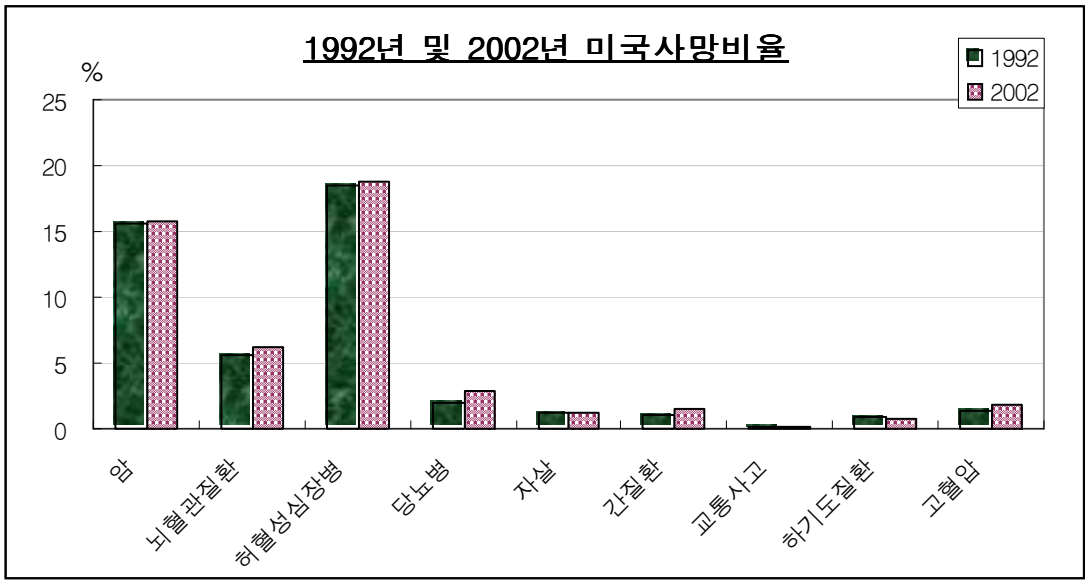
주요 10대 사망원인에 대하여 네 가지의 설명 요인별로 모두 연관성이 존재하였으며 만성간질환, 교통사고, 자살, 뇌혈관질환 및 비만이 성별에 따라 차이를 보였다. 그리고 고령을 갈수록, 결혼을 했거나 미망인의 경우, 그리고 의무교육과정까지 마친 사람일수록 주요 10대 사망원인별 사망빈도가 가장 높았다. 또한 수량화 방법을 이용하여 10대 사망원인이 연령과 가장 관련성이 높게 나타났으며 연령별로 20대 이하에서 자살 및 교통사고가 사망빈도가 높으며, 고령으로 갈수록 성인병 질환에 의한 사망이 높은 것으로 나타났다.

미국에서 특히 사망률이 높은 성인병 질환인 허혈성 심장병, 뇌혈관질환, 고혈압 및 당뇨병에 초점을 맞춰서 심도분석을 하였다. 성인병질환별로 1992년 1월부터 2002년 12월까지의 11년간 시계열 자료에 적절한 예측모형을 찾아 적용하고, 향후 2003년도의 성인병 질환에 대하여 사망빈도도 예측해 보았다. 그 결과 오차항이 AR(1)과정을 따르는 모형과 승법계절 ARIMA모형이 성인병질환에 대한 예측모형으로써 적절하게 나타났다. 오차항이 AR(1)과정을 따르는 모형에서 성인병 질환이 1월 추정계수가 가장 컸으며 9월 추정계수가 가장 작게 나타났고, 승법계절 ARIMA모형의 경우 비정상성을 가지고 있었으며 정상성을 만족하기 위한 차분을 필요하였다. 예측모형으로 적절하다고 보인 오차항이 AR(1)과정을 따르는 모형과 승법계절 ARIMA모형

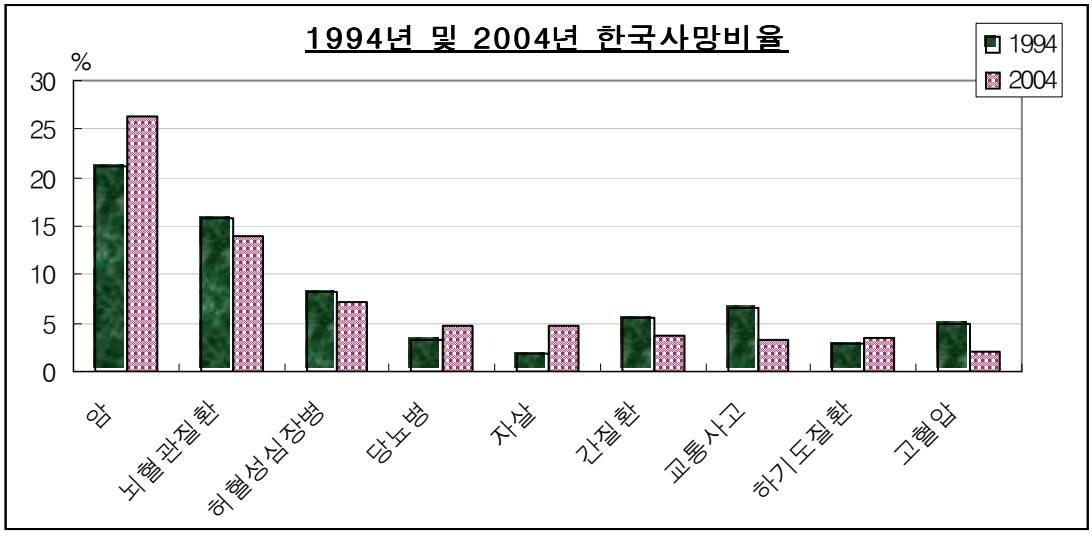
에 대하여 2003년의 12개월을 예측해 본 결과 예측빈도에는 큰 차이를 보이지 않았으며, 두 모형에 대하여 AIC와 SBC를 비교해 본 결과 오차항이 AR(1)과정을 따르는 모형보다 승법계절 ARIMA모형을 이용하는 것이 더 적절할 것으로 보여 졌다.

주요 사망 질환에 대하여 <그림 17>은 1992년과 2002년도 미국의 사망비율을 그래프로 나타낸 것이며, <그림 18>은 1994년과 2004년도 한국의 사망비율을 그래프로 나타낸 것이다.

<그림 17> 1992년 및 2002년 미국사망비율



<그림 18> 1994년 및 2004년 한국사망비율




미국의 경우 1992년과 2002년의 주요 사망원인이 차지하고 있는 백분위가 비슷한 비율을 차지하고 있으나, 한국의 경우 주요 사망원인이 차지하는 비율이 변화함을 볼 수 있다. 암, 당뇨병과 같은 미국에서 높은 비율을 가진 질환이 증가하였으며 간질환, 교통사고와 같이 낮은 비율을 보이고 있는 질환에 대해 사망비율이 줄어들었다. 즉, 한국의 사망비율이 어느 정도 미국의 사망비율과 비슷한 유형으로 변화하고 있다고 보여 진다. 따라서 우리가 미국의 자료를 이용하여 기초분석을 하고 예측모형을 알아본 결과가 한국의 사망경향을 파악하는데 있어 도움이 될 것이라 생각된다.

본 논문에서 제시한 기초분석 및 예측모형을 토대로 주요 사망원인의 기초적인 특성과 성인병 질환의 사망예측을 이용하여 의학분야나 혹은 경제분야에서 유익하게 사용될 것이라 생각된다. 그러나 우리나라 주요 사망원인에 대한 자료수집이 충분하게 되지 않아 사망원인과 잘 적용될 수 있는지에 대하여 비교를 하지 못하였다. 따라서 향후 연구에서는 한국의 사망자료를 이용하여 본 논문에서 이용한 10대 주요사망원인에 대한 기초분석을 실시하여

보고 더 나아가 성인병질환에 대한 시계열모형이 잘 적합하고 있는지에 대한 비교 분석이 필요할 것이라 생각한다.

# 참 고 문 헌

[1] 박홍현(2002). 공중보건학, 광문학.

[2] 이종협 · 최기현(2000). SAS/ETS를 이용한 시계열 분석과 그 응용, 자유아카데미.

[3] 조신섭 · 손영숙(2002). SAS/ETS를 이용한 시계열 분석, 율곡출판사.

[4] Yaffee. R.(2000). Time series analysis and forecasting, Academic Press.

[5] Wei. W.W.S.(1990). Time series analysis - Univariate and Multivariate Methods, Addison Wesley.

# ABSTRACT

## Time Series Analysis on Major Diseases' Mortality Data

Hyang-Sun Kim
Department of Statistics
The Graduate School
Sungshin Women's University

The cause of death includes various factors such as politics, economy, society and culture. Recently in Korea, a member of the organization for economic cooperation and development, there are more deaths caused by non-infectious diseases than infectious diseases. Studying causes of death in advanced countries allows Korea to prepare for changes in major causes of death.

This thesis analyses the overall trend of ten major death causes by using frequency analysis and exploratory quantification method Ⅱ. The study is based on the U.S. mortality data from January 1992 to December 2002. Regression model and multiplicative seasonal time series models are applied to statistics on adult disease deaths, such as ischemic heart disease, cerebrovascular disease, hypertension and diabetes. The optimal prediction models are chosen, which are then useful to study the patterns of adult disease deaths in Korea.

# 감사의 글

어느덧 시간이 흘러 대학원 석사과정을 마무리하는 졸업논문을 쓰게 되었습니다. 2년이라는 시간동안 많은 것을 보고 느꼈기에 대학원 생활이 더욱 소중하기만 합니다.

지난학기 내내 논문을 지도해주시고 그 이상의 많은 것을 가르쳐 주신 이종협 교수님께 감사드리며, 항상 관심을 가져주시고 지켜봐 주신 이해용 교수님, 이우선 교수님, 송일성 교수님과 논문에 많은 도움을 주신 박유성 교수님께도 진심으로 감사드립니다.

논문을 쓰면서 많이 도와 준 민정언니, 현경에게 감사하며, 힘들 때 여러 도움을 준 착한 후배 영은에게도 진심으로 감사하는 마음을 전하고자 합니다.

언제나 저를 믿어주신 부모님과 힘이 되어준 동생 해심, 영부에게 감사의 뜻을 전하며 힘들 때 응원군이 되어준 친구 설, 혜란, 승연, 그 외에 많은 고마운 분들께 감사의 뜻을 전합니다.

# 부록 A. 모형에 대한 예측값

| | | 오차항이 AR(1)과정을 따르는 모형 | | | 승법계절 ARIMA | | |
|---|---|---|---|---|---|---|---|
| | | FORECAST(y_1) | L95 | U95 | FORECAST(y_1) | L95 | U95 |
| 허혈성 심장병 | 1월 | 45664.64 | 42734.82 | 48594.46 | 46045.79 | 43046.23 | 49045.35 |
| | 2월 | 40080.57 | 36699.83 | 43461.31 | 39742.09 | 36087.20 | 43396.99 |
| | 3월 | 42486.92 | 38925.19 | 46048.64 | 41493.99 | 37560.55 | 45427.42 |
| | 4월 | 39376.88 | 35737.67 | 43016.09 | 38454.24 | 34392.66 | 42515.82 |
| | 5월 | 39101.87 | 35428.39 | 42775.35 | 38546.77 | 34424.52 | 42669.02 |
| | 6월 | 36785.78 | 33096.61 | 40474.95 | 36135.41 | 31984.07 | 40286.75 |
| | 7월 | 37184.35 | 33487.49 | 40881.20 | 36868.43 | 32703.07 | 41033.80 |
| | 8월 | 36423.57 | 32722.34 | 40124.80 | 36299.65 | 32127.51 | 40471.80 |
| | 9월 | 36121.11 | 32416.63 | 39825.59 | 36029.07 | 31853.63 | 40204.50 |
| | 10월 | 39019.86 | 35312.19 | 42727.54 | 38839.54 | 34662.52 | 43016.56 |
| | 11월 | 39334.72 | 35623.33 | 43046.12 | 39446.55 | 35268.76 | 43624.34 |
| | 12월 | 44022.39 | 40306.60 | 47738.18 | 43806.11 | 39627.94 | 47984.27 |
| 뇌혈관 질환 | 1월 | 15084.94 | 14276.97 | 15892.91 | 15110.09 | 14385.06 | 15835.12 |
| | 2월 | 13471.91 | 12545.77 | 14398.04 | 13487.29 | 12629.39 | 14345.19 |
| | 3월 | 14159.80 | 13189.55 | 15130.05 | 14180.66 | 13275.05 | 15086.28 |
| | 4월 | 13147.27 | 12159.47 | 14135.08 | 13134.03 | 12210.02 | 14058.05 |
| | 5월 | 12991.04 | 11996.01 | 13986.06 | 12978.75 | 12047.47 | 13910.03 |
| | 6월 | 12188.43 | 11190.32 | 13186.53 | 12163.04 | 11228.87 | 13097.20 |
| | 7월 | 12252.16 | 11252.62 | 13251.69 | 12235.71 | 11300.39 | 13171.03 |
| | 8월 | 12275.96 | 11275.61 | 13276.31 | 12258.23 | 11322.45 | 13194.01 |
| | 9월 | 12199.38 | 11198.38 | 13200.38 | 12183.51 | 11247.54 | 13119.47 |
| | 10월 | 13286.32 | 12284.60 | 14288.03 | 13273.56 | 12337.53 | 14209.60 |
| | 11월 | 13302.57 | 12299.97 | 14305.17 | 13284.35 | 12348.29 | 14220.42 |
| | 12월 | 14617.92 | 13614.23 | 15621.62 | 14602.20 | 13666.12 | 15538.28 |

| | | 오차항이 AR(1)과정을 따르는 모형 | | | 승법계절 ARIMA | | |
|---|---|---|---|---|---|---|---|
| | | FORECAST(y_1) | L95 | U95 | FORECAST(y_1) | L95 | U95 |
| 고혈압 | 1월 | 4528.60 | 4280.87 | 4776.32 | 4523.84 | 4307.95 | 4739.73 |
| | 2월 | 3999.21 | 3716.52 | 4281.90 | 4016.54 | 3776.06 | 4257.02 |
| | 3월 | 4166.55 | 3871.41 | 4461.69 | 4210.63 | 3959.93 | 4461.33 |
| | 4월 | 3873.42 | 3573.56 | 4173.28 | 3932.51 | 3674.81 | 4190.22 |
| | 5월 | 3792.14 | 3490.44 | 4093.84 | 3861.06 | 3597.31 | 4124.80 |
| | 6월 | 3637.06 | 3334.61 | 3939.51 | 3714.72 | 3445.31 | 3984.13 |
| | 7월 | 3701.69 | 3398.90 | 4004.47 | 3784.69 | 3509.80 | 4059.59 |
| | 8월 | 3633.96 | 3330.98 | 3936.93 | 3720.91 | 3440.67 | 4001.16 |
| | 9월 | 3584.65 | 3281.51 | 3887.79 | 3673.64 | 3388.15 | 3959.13 |
| | 10월 | 3819.22 | 3515.89 | 4122.54 | 3911.38 | 3620.75 | 4202.02 |
| | 11월 | 3871.07 | 3567.49 | 4174.65 | 3967.20 | 3671.51 | 4262.89 |
| | 12월 | 4261.80 | 3957.91 | 4565.70 | 4363.04 | 4062.38 | 4663.71 |
| 당뇨병 | 1월 | 6694.14 | 6389.79 | 6998.49 | 6694.33 | 6402.83 | 6985.83 |
| | 2월 | 6016.97 | 5677.69 | 6356.25 | 5979.66 | 5653.14 | 6306.18 |
| | 3월 | 6305.29 | 5956.29 | 6654.29 | 6273.53 | 5935.99 | 6611.07 |
| | 4월 | 5953.71 | 5601.84 | 6305.58 | 5876.29 | 5533.58 | 6219.01 |
| | 5월 | 5936.05 | 5583.29 | 6288.80 | 5850.71 | 5504.63 | 6196.78 |
| | 6월 | 5620.87 | 5267.83 | 5973.91 | 5514.31 | 5165.53 | 5863.08 |
| | 7월 | 5702.16 | 5349.00 | 6055.32 | 5587.99 | 5236.78 | 5939.20 |
| | 8월 | 5645.94 | 5292.71 | 5999.18 | 5522.71 | 5169.18 | 5876.24 |
| | 9월 | 5601.74 | 5248.42 | 5955.06 | 5493.14 | 5137.34 | 5848.93 |
| | 10월 | 5941.41 | 5587.95 | 6294.87 | 5818.91 | 5460.88 | 6176.93 |
| | 11월 | 6003.73 | 5650.05 | 6357.41 | 5912.35 | 5552.12 | 6272.59 |
| | 12월 | 6536.67 | 6182.69 | 6890.65 | 6466.13 | 6103.70 | 6828.56 |

# 부록 B. 기초분석 프로그램

```
libname time 'C:₩mort₩mort02';
data time.time02;
infile 'C:₩mort₩mort02₩MORT02.dat';
input  gen_a $ 115-118 gen_b $ 3 gen_c $ 19 gen_d $ 20 gen_e $ 55-56 gen_f $ 83
       dec_a $ 59 dec_b $ 62 dec_c1 $64-66 dec_c2 $ 69-70 dec_d $ 77 dec_e $ 78-79 dec_f
$ 82 dec_g $ 85-87 dec_h $ 88-90 dec_i $ 54 cause $ 146-148;run;

data group_1; set time.time02;length grp $6.;
if cause in('072','073','074','075') then grp='a1';if cause='077' then grp='a2';
if cause='078' then grp='a3';if cause='081' then grp='a4';
if cause in('085','086') then grp='a5';if cause='088' then grp='a6';
if cause='092' then grp='a7';if cause='093' then grp='a8';
if cause='104' then grp='a9';if cause='107' then grp='a10';
if cause='113' then grp='a11';if cause='118' then grp='a12';
if cause='122' then grp='a13';if cause in('134','135','136','137') then grp='a14';
if cause='159' then grp='b';
if cause in('206','207','208','209') then grp='c';
if cause in('211','212','213','214','215') then grp='d';
if cause in('235','236','237','238','239') then grp='e';
if cause in('264','265','266','268') then grp='f';
if cause in('298','299','300','301','302') then grp='g';
if cause='386' then grp='h';
if cause in('425','426','427','428','429','430','431') then grp='i';
if cause='169' then grp='j';run;

data group_2;set  group_1;length agegrp $4. age_g $4.;
if grp='' then delete;
if  grp  in('a1','a2','a3','a4','a5','a6','a7','a8','a9','a10','a11','a12','a13','a14')
then dis='a'; else dis=grp;
if dec_c2 in('01','02','03','04','05','06','07') then agegrp='1';
if dec_c2 in('08','09') then agegrp='2';if dec_c2 in('10','11') then agegrp='3';
if dec_c2 in('12','13') then agegrp='4';if dec_c2 in('14','15') then agegrp='5';
if dec_c2 in('16','17') then agegrp='6';if dec_c2 in('18','19') then agegrp='7';
if dec_c2 in('20','21') then agegrp='8';if dec_c2 in('22','23') then agegrp='9';
if dec_c2 in('24','25') then agegrp='10';if dec_c2='26' then agegrp='11';
age_g=agegrp;if agegrp in('1','2') then age_g='1';if agegrp in('9','10','11')
    then age_g='9';
```

```
label dis='10개 병명' gen_e='사망월' dec_a='성별' dec_b='인종' dec_d='결혼유무' dec_i='
교육기간';run;

data grp;set group_2;if agegrp='' then delete;if grp='a9' and dec_a='1' then delete;
m=substr(dec_c1,1,1);if m in('0','1') then age1=substr(dec_c1,2,2);else if m='9'
  then delete; else age1='0';age2=age1*1;
if dec_i='6' then delete;if dec_d in('8','9') then delete;if age2>5;run;
/*수량화2의 방법이용-10대병명*/
data b; set grp;
if dis='a' then dis1=1;else dis1=0;if dis='b' then dis2=1;else dis2=0;
if dis='c' then dis3=1;else dis3=0;if dis='d' then dis4=1;else dis4=0;
if dis='e' then dis5=1;else dis5=0;if dis='f' then dis6=1;else dis6=0;
if dis='g' then dis7=1;else dis7=0;if dis='h' then dis8=1;else dis8=0;
if dis='i' then dis9=1;else dis9=0;if dec_a='1' then sex=1;else sex=0;
if dec_i='1' then edu1=1;else edu1=0;if dec_i='2' then edu2=1;else edu2=0;
if dec_i='3' then edu3=1;else edu3=0;if dec_i='4' then edu4=1;else edu4=0;
if dec_d='1' then mar1=1;else mar1=0;if dec_d='2' then mar2=1;else mar2=0;
if dec_d='3' then mar3=1;else mar3=0;if age_g='1' then age_1=1;else age_1=0;
if age_g='3' then age_3=1;else age_3=0;if age_g='4' then age_4=1;else age_4=0;
if age_g='5' then age_5=1;else age_5=0;if age_g='6' then age_6=1;else age_6=0;
if age_g='7' then age_7=1;else age_7=0;if age_g='8' then age_8=1;else age_8=0;
run;

proc cancorr data=b vname='disease' wname='factor' ;
var dis1 dis2 dis3 dis4 dis5 dis6 dis7  dis8  dis9;
with sex edu1 edu2 edu3 edu4 age_1 age_3 age_4 age_5 age_6 age_7 age_8
      mar1 mar2 mar3;run;
```

# 부록 C. 허혈성 심장병 자료에 대한 시계열 분석 프로그램

```
/*허혈성심장병에 대한 시계열분석*/
data d;
input zt @@;
t=_n_;
cards;
39709 33662 35048 33134 33039 31262 31247 31155 30701 33357 33393 37281
37448 34611 39429 34884 33938 31705 31980 31065 30992 33701 34218 39476
44941 36067 37588 35392 35673 33203 33174 32682 32505 35477 34799 38696
40397 36568 39195 35974 35690 33103 34039 32800 32346 34488 36344 40523
41885 36354 38166 35642 34988 32292 32955 31900 31593 35171 35301 41906
42579 35141 37869 36015 35166 32615 32454 32233 31814 34563 35546 40081
42951 37182 37829 34487 34803 33048 32490 31882 31257 34109 34293 37267
48472 43165 46459 40157 38720 36120 37496 35525 35689 39231 38837 46481
50892 40671 40345 37988 38115 35454 36329 35991 36192 38645 39578 44539
45145 39138 42260 39360 39031 36735 36352 35508 35010 38386 37035 40980
44041 40432 42601 37783 37469 34862 35783 34893 34048 36854 38150 41985
;run;
proc gplot data=d;
plot zt*t; symbol i=join v=dot;
run;

/*ARIMA모형*/
proc arima data=d;identify var=zt nlag=40;/*identify var=zt(1) nlag=50;*/
identify var=zt(12) nlag=40;
estimate p=1 q=(1,2)(12) method=ml noconstant plot ;
forecast lead=12 out=adult_d;
run;

/*일반 선형회귀모형*/
data d_1;set d;m=mod(t,12);
if m=1 then t_1=1;else t_1=0;if m=2 then t_2=1;else t_2=0;if m=3 then t_3=1;else
t_3=0;if m=4 then t_4=1;else t_4=0;
if m=5 then t_5=1;else t_5=0;if m=6 then t_6=1;else t_6=0;if m=7 then t_7=1;else
t_7=0;if m=8 then t_8=1;else t_8=0;
if m=9 then t_9=1;else t_9=0;if m=10 then t_10=1;else t_10=0;if m=11 then t_11=1;else
t_11=0;if m=0 then t_12=1;else t_12=0;run;
```

```
proc reg data=d_1 ;
model zt=t t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_10 t_11  t_12/DW  noint;
output out=res r=residual p=pred;run;
proc gplot data=res ;plot residual*t/vref=0;run;



/*자기회귀오차모형*/
proc autoreg data=d_1 outest=v1;
model zt=t t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_10 t_11  t_12/  noint partial nlag=10;
model zt=t t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_10 t_11  t_12/  noint partial nlag=1
method=ml;
output out=d_2 r=r_1 p=y_1 LCL=L95 UCL=U95;run;

proc gplot data=d_2 ;
plot y_1*t;run;

/*예측 data 만들기*/
data d_3;zt=.;do t=133 to 144;output;end;run;

data fore_d;merge d_1 d_3;by t;m=mod(t,12);
if m=1 then t_1=1;else t_1=0;if m=2 then t_2=1;else t_2=0;if m=3 then t_3=1;else
t_3=0;if m=4 then t_4=1;else t_4=0;
if m=5 then t_5=1;else t_5=0;if m=6 then t_6=1;else t_6=0;if m=7 then t_7=1;else
t_7=0;if m=8 then t_8=1;else t_8=0;
if m=9 then t_9=1;else t_9=0;if m=10 then t_10=1;else t_10=0;if m=11 then t_11=1;else
t_11=0;if m=0 then t_12=1;else t_12=0;run;

proc autoreg data=fore_d ;
model zt=t t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_10 t_11  t_12/  noint partial nlag=1
method=ml;
output out=d_5 r=r_1 p=y_1 LCL=L95 UCL=U95;run;
```