Annals of Statistical Research
Vol. 4. No. 1, 2008, pp.1~14

A Study on the overall evaluation on the 7th curriculum implementation training of middle school teachers after the training finished[1)]

Soo-dong Kim[2)]

Abstract

Through summative evaluation on the 7th curriculum implementation training of 48 middle school teachers as a survey of questionnaire which is practiced on training participants, overall satisfaction levels & hereafter improvement strategies after training are analysed. That is, first, the background characteristics of the training participants are analysed. Second, the satisfaction levels & improvement strategies on training text, lecture of the lecturer & training program are analysed. Lastly, some strategies for improvement of the training are suggested as conclusion.

Key Words: 7th national curriculum, 7th curriculum implementation training of middle school teachers, training program development for curriculum implementation of middle school teachers, curriculum implementation training system, curriculum implementation training method

---

1) 이 연구는 「조덕주, 김수동 (2000). 중학교 교육과정·평가 연수 프로그램 개발 연구. 한국교육과정평가원」의 '2000년 중학교 연수 프로그램 실시 및 평가'의 일부 내용을 수정, 보완한 것임.

2) Associate Professor, College of Education, 707, Seokjang-dong, Gyeongju, Gyeongsangbuk-do, 780-714 Korea. E-mail: lyskhj1201@dongguk.ac.kr

Ⅰ. 서론

2000년도 하계 방학 중에 한국교육과정평가원은 전국 16개 시·도교육청에서 추천한 중학교 교사들(교원과 전문직 포함)을 대상으로 제7차 교육과정 운영에 대한 연수를 실시한 이후, 연수 전반에 대한 총합평가를 실시하였으며, 그 결과를 토대로 향후 연수 개선을 위한 시사점을 도출하였다. 먼저 교육과정과 연수 일정은 다음의 <표 Ⅰ-1>과 같다.

<표 Ⅰ-1> 연수 일정

시 간	7월 18일(화)	7월 19일(수)	7월 20일(목)	7월 21일(금)	7월 22일(토)
1 9:00 - 12:00	개강식 및 연수를 위한 orientation	재량활동에 대한 정의, 유형, 운영 방안(이론 및 실습)	학교교육과정 작성의 실제 -편제, 시간배당 -교과, 특활, 생활지도	수행평가에 대한 이해 및 오해	학습부진아 지도의 이론과 실제 (9:00-13:00)
중식 12:00 - 13:00					
2 13:00 - 16:00	수준별 교육과정의 정의 및 운영 방안	재량활동 사례발표 및, 토의 (13:00-18:00)	학교교육과정 작성의 실제 -지원활동 -부진아지도 -평가 계획 (13:00-18:00)	성취기준과 평가 기준의 작성	
시 간	7월 24일(월)	7월 25일(화)	7월 26일(수)	7월 27일(목)	7월 28일(금)
1 9:00 - 12:00	제7차 국어과 교육과정의 이해	국어과 성취 기준의 개발과 활용 Ⅰ	국어과 학습 평가 도구의 개발과 활용 Ⅰ	국어과 수준별 교육과정 구현에 적합한 교수·학습 자료 Ⅰ	평가 및 수료식
	제7차 영어과 교육과정의 이해	영어과 성취 기준 및 평가 기준의 개발과 활용 Ⅰ	영어과 평가의 실제 Ⅰ	영어과 활동 중심의 교수·학습 방안 Ⅰ	
	제7차 수학과 교육과정의 이해	수학과 성취 기준 및 평가 기준의 개발과 활용 Ⅰ	수학과 평가의 실제 Ⅰ	수학과 멀티미디어 를 이용한 교수· 학습 방안 Ⅰ	
중식 12:00 - 13:00					
2 13:00 - 16:00	국어과 수준별 교육과정의 운영방안	국어과 성취 기준의 개발과 활용 Ⅱ	국어과 학습 평가 도구의 개발과 활용 Ⅱ	국어과 수준별 교육과정 구현에 적합한 교수· 학습 자료 Ⅱ	귀 가
	영어과 수준별 교육과정의 운영방안	영어과 성취 기준 및 평가기준의 개발과 활용 Ⅱ	영어과 평가의 실제 Ⅱ	영어과 활동 중심의 교수·학습 방안 Ⅱ	
	수학과 수준별 교육과정의 운영 방안	수학과 성취 기준 및 평가기준의 개발과 활용 Ⅱ	수학과 평가의 실제 Ⅱ	수학과 멀티미디어 를 이용한 교수· 학습 방안 Ⅱ	
3 17:00 - 20:00	분임토의 : 학교교육과정 작성 (17:00-21:00)	분임토의: 학교교육과정 작성			

위의 <표 Ⅰ-1>에서 알 수 있는 바와 같이 교육과정은 기초 영역, 국어, 수학 및 영어 영역, 분임 토의 영역의 60시간 4학점으로 구성되어 있다. 강의는 기초 영역 6강좌,

국어, 수학, 영어 영역의 각각 5강좌로 모두 21강좌이지만 교과 영역에서 1강좌에 2명의 강사가 담당하는 경우도 있으므로 강의는 모두 23개이다.

1. 평가 목적

연수 실시 후 평가의 구체적 목적은 다음과 같다.

첫째, 연수 대상자들의 배경특성은 무엇인지 알아본다.

둘째, 연수 대상자들의 연수교재, 강사의 강의, 연수 프로그램에 대한 만족도와 개선의견을 알아본다.

셋째, 연수의 전반적인 평가 결과를 토대로 향후 연수의 개선 방안을 위한 결론을 도출한다.

2. 평가 방법

연수 실시 후 참여한 연수 대상자들에게 설문조사 형태의 '연수에 대한 총합 평가'를 실시하여 연수에 관한 전반적인 만족도와 향후 개선 방안을 살펴보았다. 먼저 연수 대상자들의 연령 및 직책, 성별 및 학력, 연수의 경험, 연수를 받는 목적 등과 같은 배경특성을 알아보았다.

그리고 연수 대상자들의 연수교재, 강사의 강의, 연수 프로그램에 대한 만족도와 개선에 대한 의견을 알아보았다. 그리고 마지막으로 연수 개선을 위한 방안을 결론으로 도출하였다.

3. 평가 내용

설정된 평가영역은 연수교재, 강사의 강의, 연수 프로그램에 관한 것이다. 각 영역에 대한 구체적인 항목들은 다음 <표 Ⅰ-2>와 같다.

<표 Ⅰ-2> 평가 영역 및 평가 항목

평가 영역	평가 항목
연수교재	· 연수교재 내용의 적절성 · 연수교재 내용 이해의 용이성 · 구성방식과 예시 · 교재의 유용성
연수 강의 및 강사	· 강의 내용의 적절성 · 강사의 피드백 · 강사의 수업진행방식
연수프로그램 전반	· 연수 전체에 대한 만족도 · 주제와 관련된 연수 내용의 적절성 · 연수 이해의 정도 · 실제와의 관련성

Ⅱ. 7차 중학교 교육과정 운영 연수 프로그램 평가

1. 연수 대상자들의 배경특성

가. 연령 및 직책

2000년에 연수를 받은 48명의 중학교 교사들의 연령층은 30대부터 50~60세까지 비교적 다양하였으나 주로 30대와 40대가 대부분을 차지하였고, 결과는 다음 <표 Ⅱ-1>에서 보는 바와 같다.

<표 Ⅱ-1> 연수 대상자들의 연령

연령 구분 / 분포	30세이하	30~40세	40~50세	50~60세
분포(%)	0 %	28.6%	54.8 %	16.7 %

한편, 연수 대상자들의 직책은 평교사가 31.3%, 부장교사가 50%를 차지하여 전체의 81.3%를 점하였으며, 세부 결과는 다음 <표 Ⅱ-2>와 같다.

<표 Ⅱ-2> 연수 대상자들의 직책

직책 / 분포	부장교사	교사	장학사, 연구사
분포 (%)	50 %	31 %	19 %

나. 성별, 학력 및 교직경력

성비는 남, 여교사가 78.6%와 21.4%의 비율을 차지하였고, 다음 <표 Ⅱ-3>에서 보는 바와 같이, 연수 대상자들의 학력은 대학원 졸업학력자가 59.5%로 매우 높은 편이었고, 대학원 과정중인 교사도 23.8%를 차지하였다.

<표 Ⅱ-3> 연수 대상자들의 학력

학력 / 분포	대학 졸업	대학원과정 중	대학원 졸
분포(%)	16.7 %	23.8 %	59.5 %

다. 연수 신청의 이유

이번 연수를 신청한 가장 중요한 이유로는 다음 <표 Ⅱ-4>에서 보는 바와 같이, 전문성 신장이 절반 이상으로 76.4%를 차지하였으며, 그 다음은 학점이나 수료증 취득으로 12.5%, 다음은 시·도 교육청이나 학교장의 권유로 6.3%, 다음으로 주위 사람들의 권유로 4.8% 순으로 나타났다. 중학교 연수 대상자들은 전공이 국어, 수학, 영어 교과인 만큼 교육 현장에서의 전문성과 실용성 요구와 부합되어 제7차 교육과정을 통하여 국어과 심화보충형, 수학과 및 영어과의 단계형 수준별 수업 운영에 관한 내용을 구체적으로 이해하고 학교현장에 적용하기를 적극적으로 희망하였다.

<표 Ⅱ-4> 한국교육과정평가원 연수를 신청한 이유

항목 / 중요순서	전문성 신장	학점이나 수료증 취득	시도교육청이나 학교장권유	주위 사람들의 권유
분포(%)	76.4 %	12.5 %	6.3 %	4.8%

2. 연수교재, 연수강의와 강사 및 연수 프로그램 전반에 대한 만족도에 대한 의견

먼저 연수 실시 후 연수교재, 연수강의와 강사, 연수 프로그램 전반에 대한 만족도를 분석하고, 연수 대상자들이 답변한 자유-기술식 응답결과를 토대로 개선에 대한 의견을 조사하였다.

가. 연수교재

1) 만족도 분석

2000년에 실시한 제7차 중학교 교육과정 운영 연수는 '제7차 교육과정'에 대한 기초 분야 6개 강의와 국어, 수학 및 영어교과 5~6개 강의로 총 23개 강의가 개설되었다. 각 강의마다 연수자료가 개발되었으며 이러한 연수교재에 대한 종합적인 평가는 다음 <표 Ⅱ-5>에서 보는 바와 같이, 연수 교재 내용 구성의 적절성, 이해의 용이성, 구성방식과 예시, 유용성이란 4개의 측면으로 이루어 졌으며, 5점 만점의 척도로 평정되었다.

교재에 대한 전체적인 평가는 4.25점으로 연수교재 주제와 내용의 적절성, 이해의 정도와 구성방식과 예시, 그리고 자료로서의 유용성 등이 대체적으로 만족함을 알 수 있었다.

<표 Ⅱ-5> 연수교재의 만족도

평가항목 / 강의명	내용구성	이해정도	구성방식과 예시	유용성
평 균	4.3	4.2	4.2	4.3

2) 만족도 개선을 위한 시사점

연수 교재의 개선에 대한 의견을 자유 진술한 결과, '만족 한다'와 '이론중심의 강의 자료'가 일부 의견으로 나타났다. 이론 중심적인 강의 자료였다는 응답결과로 미루어 보아, 연수 대상자들이 이론중심의 강의 자료나 내용보다는 현장에서의 실제적으로 적용 가능한 프로그램과 자료를 제공받기 원하는 것을 알 수 있다. 이와 같은 문제점은 여러 번 제기되어 왔으므로 관심을 갖고 개선되어야 할 사항으로 생각된다. 그 다음으로 '현장사례가 부족한 연수 자료의 문제점'을 일부 지적하였다. 이는 앞에서 개선점으로 가장 많이 지적된 이론중심의 강의 자료와 맥을 같이 하는 응답이라 할 수 있겠다. 즉, 원론적인 내용보다는 현재 학교에서 실제적으로 시행되고 있는 모범 사례 소개를 실례로 들어 좀 더 교육현장에 쉽게 적용할 수 있기를 바라고 있음을 알 수 있다.

나. 연수강사의 강의

1) 만족도 분석

23개의 연수 강의별로 강사의 강의에 대한 평가는 다음 <표 Ⅱ-6>에서 보는 바와 같이, 5점 만점 기준으로 최하 평균점수 3.6부터 최고점수 5.0의 점수분포를 보였고, 3개의 하위 평가 영역인 강의내용, 강사의 피드백, 수업진행방식의 만족도에서는 평균 4.18의 점수를 받음으로써 2000년의 연수강사들의 강의 내용이나 강의 진행방식이 적절하였음을 보여주고 있다.

<표 Ⅱ-6> 강사의 강의에 대한 만족도

항목 / 강의명	강의 내용	강사의 피드백	수업진행방식	전체적 평가
	평균	평균	평균	평균
1	4.1	4.7	4.0	4.3
2	4.3	4.2	4.1	4.2
3	4.2	4.4	4.3	4.3
4	4.0	3.8	3.9	3.9
5	4.4	4.3	4.3	4.3
6	4.5	4.4	4.4	4.4
7	4.4	4.7	4.6	4.6
8	4.0	4.2	4.1	4.1
9	4.4	4.6	4.6	4.5
10	4.6	4.5	4.6	4.6
11	4.6	4.6	4.7	4.6
12	5.0	5.0	5.0	5.0
13	3.3	2.9	2.9	3.0
14	4.0	3.8	3.8	3.9
15	4.5	4.4	4.3	4.4
16	4.5	4.5	4.7	4.6
17	4.2	4.1	4.0	4.1
18	3.7	3.9	3.5	3.7
19	3.8	3.6	3.7	3.7
20	3.9	3.4	3.5	3.6
21	4.3	4.1	4.2	4.2
22	4.3	4.1	4.0	4.1
23	4.1	3.8	4.0	4.0
전체 평균	4.22	4.17	4.14	4.18

2) 만족도 개선을 위한 시사점

자유-서술식 응답결과에서 강사의 강의에 대한 개선점을 분석하면 일부의 응답자가 강사의 태도에 불만족을 나타냈는데 예를 들어 '질문에 대해 성심 성의껏 답해 주기를 바란다', '본인의 의지만을 너무 피력하신 것 같다' 등의 의견이 나왔다. 그리고 강의 스타일에 대한 불만족, 이론중심의 강의에 대한 개선촉구, 강의 중 사례 소개나 연수생들 간의 자유로운 토의 부족이 나타났다.

강의와 관련하여 자유-기술한 응답 중에 '강의에 만족 한다'고 일부 답하였는데 어떤 점이 좋았는지 분석하여 보면 다음과 같다. 강의에 만족하는 이유로는 사례를 통한 수업이 만족스럽다는 점이다. 특히 현장 교사의 사례를 통한 수업이 좋았다는 응답이 많았고, 현장에서 활용될 수 있는 다양한 자료를 통한 수업이 만족스러웠다라고 답했으며, 또한 현장에서 적용할 수 있는 실례를 제시해 좋았다는 의견 등이 있었다. 또 다른 이유로는 '제7차 교육과정에 대한 의문점이 해소되었다', '학자적인 분위기로 차분하게 전달해주신 강의 전달내용으로 인해 제7차 교육과정에 대해 알 수 있어 좋았다', '전문적이고 깊이 있는 자세로 확실한 답변을 해 주셔서 좋았다' 등 강의 내용에 대한 여러 의견이 있었다.

강사에 대해서는 7차 교육과정의 전문가인 강사들의 강의가 좋았으며 다른 한편으로는 실질적인 현장의 상황을 연구하는 연구학교 현직 교사들의 강의도 좋았음을 지적하고 있다. 따라서 연수 강사의 구성을 교육과정 및 교육평가 전문가와 실질적인 학교 현장의 전문가로 혼합 구성하는 것이 바람직하다고 보여 진다.

다. 연수 프로그램

1) 만족도 분석

연수 프로그램의 만족도는 다음 <표 Ⅱ-7>에서 보는 바와 같이, 5점 만점을 기준으로 전체 평균 4.2로 높은 것으로 나타났다. 전반적으로 제7차 중학교 교육과정 운영에 대한 이해도를 높이고 학교 현장에 이를 적용시키는데 도움이 될 것으로 생각하는 것으로 해석할 수 있다.

<표 Ⅱ-7> 연수 프로그램의 만족도

연수명	항 목	평 균
7차 중학교 교육과정	1. “7차 교육과정” 에 관한 이번 연수는 만족스러웠다.	4.3
	2. 7차 교육과정을 이해하기 위한 연수내용을 적절하였다.	4.0
	3. 이번에 7차 교육과정에 관련된 연수를 받고 난 후, 7차 교육과정과 관련된 내용을 충분히 이해하였다고 생각한다.	4.1
	4. 이번에 7차 교육과정과 관련된 연수를 받고 난 후, 학교에서 7차 교육과정에 의한 수업을 실행하는데 큰 도움을 줄 것 같다.	4.1
	5. 이번에 7차 교육과정과 관련된 연수를 받고 난 후, 학교에서 7차 교육과정에 의한 평가를 실행하는데 큰 도움을 줄 것 같다.	4.4

2) 만족도 개선을 위한 시사점

연수 프로그램 개선에 대한 자유-서술식 응답 결과는 ‘이론중심의 수업’과 ‘강의자료의 비활용성’이 일부 나타났고, 강사 선정에 불만족이 일부 있었다.

한편 이론중심 수업의 대안 책으로 제시한 내용으로는 교사들에게 현실적으로 필요한 교수방법이나, 현장교사의 사례소개나 토의를 통한 수업, 혹은 학습활동에서 그룹토론을 많이 하자는 의견이 나왔다. 또 강의 자료의 비활용성을 지적했는데 일부의 응답자가 원론적인 이론보다는 현장에서 적용 가능한 프로그램과 자료를 제공했으면 하는 의견이 나왔다. 또한 연수교사들에게 상호 교류할 기회를 제공하는 것이 필요하다는 의견과 이론중심의 강의나 딱딱한 수업 전개로 강사에 대해 만족하지 못한다는 대답도 일부 나왔다.

Ⅲ. 결론

앞의 Ⅰ, Ⅱ장에서는 연수 실시 후 연수 전반에 관한 총합평가를 통해, 연수교재, 연수강사의 강의, 연수 프로그램에 대한 만족도와 개선을 위한 시사점을 분석하였다. 이러한 결과를 토대로 향후 연수 전반의 개선을 위한 방안을 고찰해 보고자 한다.

1. 연수교재와 관련한 결론

연수교재와 관련한 결론으로는 **첫째,** 이론중심의 강의 자료와 내용보다는 현장에서 실제적으로 활용-가능한 프로그램과 자료를 보강할 필요가 있다는 것이다.

둘째, 학교현장과 상황을 고려하여 계속 업데이트되는 연수교재의 개발이 촉구된다. 연수대상자들이 연수를 받는 목적으로 전문지식 함양과 학교교육에서의 문제를 해결할 수 있는 능력을 발전시키는 것을 가장 중요하게 언급한 것에서 알 수 있듯이, 학교현장의 변화와 더불어 연수대상자들의 요구와 기대치는 날로 증가한다. 그러므로 이에 대응하여 보다 질 높은 연수교재 개발이 계속적으로 요구된다.

2. 강사의 강의와 관련한 결론

강사의 강의와 관련한 결론으로는 **첫째,** 실제 교육현장에서 적용할 수 있는 현장감 있는 강의를 통하여 이론과 실제의 조화가 요구된다. 이러한 점에서 전문적인 이론적 배경을 가지고 연구하는 강사들뿐 아니라 실질적으로 현장에서 연구하는 학교 현직 교사들을 각 주제에 알맞게 활용하는 것이 중요하다. 따라서 연수강사의 구성은 교육과정 및 교육평가 전문가와 실질적인 학교 현장의 전문가로 혼합 구성하는 것이 바람직하다고 본다.

둘째, 자유로운 토론의 장을 마련하여 활발한 질의응답, 피드백 제시를 수반하는 강의진행방식이 보다 많이 요구된다. 연수대상자들은 현직 교사나 교육전문직 종사자를 대상으로 하므로 기본적인 이론적 배경을 가지고 있을 것이다. 연수대상자들이 연수를 받는 목적에서 교사들 간의 정보교환과 의사전달능력 함양, 인간관계 개선 등을 지적한 것을 통하여 알 수 있듯이, 상호작용을 활발히 하여 정보를 교환하며 자유로운 토론의 기회를 제공하는 강의가 요구된다.

3. 연수 프로그램과 관련한 결론

연수 프로그램과 관련한 결론으로는 **첫째,** 실제현장에 가까운 이론의 소개와 함께 실제적으로 활용할 수 있는 활용 방법을 연수함이 바람직하다고 본다. 즉, 학교 현장에 시급한 것이 바로 적용할 수 있는 구체적인 방안이라고 하더라도 그 기본 배경을 이해하고 있지 못하면 그 방법이 장기적으로 그리고 창의적으로 변환되어 적용될 수 없다는 점을 생각할 때, 현장 적용방안의 이론적 배경을 연수하는 것이 필요한 일이라고 생각된다. 다른 한편으로 아무리 훌륭한 이론이 있더라도 그것이 현장에 별로 도움을 줄 수 없다면 그 효과가 줄어들고 의미가 약화될 것이다.

둘째, 연수 중 연수대상자들 간의 활발한 교류를 통해 정보교환과 친목도모를 할 수 있는 기회를 제공하는 것이 필요하다. 전국 각지에서 모이는 같은 전공자들 간의 교류 기회는 그리 흔하지 않다고 볼 때, 같은 목적을 추구하는 만남의 장을 마련하는 기회가 연수를 통해서 이루어진다면, 연수의 효율성과 효과성은 극대화될 것이다.

참고문헌

[1] 김수동, 박순경, 유승연, 권재기(2001). 고등학교 교육과정·평가 연수 프로그램 개발 연구. 한국교육과정평가원. 연구보고 RRC 2001-11-1.

[2] 이윤식, 최상근, 허병기(1994). 교사양성체제 개선 방안 연구. 한국교육개발원. RR 94-18.

[3] 조덕주(1999). '연계자를 매개로 한 교육과정 적용'에 있어서의 한계-교원연수를 중심으로. 한국교육과정평가원 Forum 자료.

[4] 조덕주, 김수동(2000). 중학교 교육과정·평가 연수 프로그램 개발 연구. 한국교육과정평가원. 연구보고 RRC 2000-2.

[5] 조덕주, 김수동, 채선희(1998). 연수 프로그램 개발 연구- 초등학교 교육과정 및 평가 방안을 중심으로 -. 한국교육과정평가원. 연구보고 RRC 98-3.

Annals of Statistical Research
Vol. 4. No. 1, 2008, pp.15~34

A Study on the prediction of ozone concentration in Seoul

Joonkeun Yum[1], Eunyoung Kim[2], Hyunjung Noh[3]

Abstract

To prevent global warming, ozone concentration prediction models including air pollutants and meteorological factor are developed. This study is carried out to access the property of Ozone and air pollutant, to analyze the correlation with ozone and others such as SO_2, NO_2, PM_{10}, CO, rainfall, wind speed, sunshine duration, humidity and temperature. The ozone concentration model is developed during the period of three years (Jan.2005~ Dec.2007) in Seoul. One is linear multiple regression model using the result of correlation with ozone. Another is ridge regression model.

Key words : ozone, multiple regression model. ridge regression model

1) Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea
2) Master's course, Department of Statistics, Dongguk University, Seoul 100-715, Korea
3) Master's course, Department of Statistics, Dongguk University, Seoul 100-715, Korea

Ⅰ. 서 론

최근 전 세계적으로 지구 연평균 기온이 상대적으로 단기간에 급격히 상승하는 지구 온난화 현상이 나타나고 있다. 지구 온난화 현상이란 자동차의 보급과 운행량의 증가로 인해 이산화질소와 오존 농도가 점점 증가하면서 열이 빠져 나가지 못해 지구 온도가 점차 높아지고 있는 것을 말한다. 지구온난화 방지를 위해 선진국은 1992년 브라질 리우에서 기후변화 협약 체결 이후 1997년 교토의정서 채택에 따라 온실가스 배출량을 1990년 수준으로 감축하고자 노력하고 있다. 우리나라는 2002년 교토의정서에 비준하면서 개발도상국 지위를 확보해 2012년까지는 온실가스 감축의무가 없었으나 2013년 의무감축국가로 지정되었다. 이에 따라 온실가스 증가율이 매년 5%에 달하는 우리나라로서는 온난화 현상으로 인한 에너지 소비를 줄이기 위해 국가적 노력이 필요하다.

온실 가스의 성분 중 하나인 오존은 많은 인구와 차량, 산업시설 등으로부터 배출된 질소산화물(NO_X), 일산화탄소(CO) 등이 대기 중에서 기상인자인 기온, 일사량, 풍속, 습도, 대기안정도, 역전층 고도 등의 영향을 받아 광화학 반응으로 생성되는 광산화 물질로써 2차 오염물질이라고도 한다.(윤오섭, 1999). 또한, 오존은 광화학 스모그의 원인이 되는 강한 산화력을 가진 물질로서 높은 농도에 장시간 노출되면 점막의 자극, 호흡기 질환 등 인간의 건강과 농작물에도 심각한 영향을 미치며, 이산화탄소보다도 훨씬 강력한 온실효과를 일으킨다고 알려져 있기도 한다.(전병일 외, 1995; 서의훈 외, 2001).

현 시점에서 오존 농도의 예측은 지구온난화 방지를 위한 오존 농도 감소에 대한 대책 수립 못지않게 중요한 과제로 떠오르고 있다. 본 연구의 목적은 대기오염 물질 자료(CO, NO_2, PM_{10}, SO_2)와 기후인자 자료(기온, 강수량, 풍속, 습도, 일조시간)를 가지고 오존농도를 예측하는 모형을 구축하고 앞으로의 오존 농도 예측 모델을 개발하는데 유용한 기초자료를 제공하는 것이다.

본 연구에서는 서울특별시의 2005년부터 2007년까지의 일평균 대기오염 물질 자료와 기상자료를 이용하여 오존과 대기 오염 물질의 시계열적 특성에 대해 알아보고, 오존과 다른 요인들 간의 상관 분석을 하였다. 오존 농도는 계절에 따라 차이를 나타내므로 계절별로 영향을 주는 변수가 다를 것으로 판단, 계절별로 상관 분석을 한 뒤 그 결과를 가지고 일차적으로 오존에 영향을 주는 변수들을 선택하여 다중 회귀 모형과 능형 회귀 모형을 구축하였다.

Ⅱ. 연구방법

1. 연구 대상 및 분석자료

본 연구의 대상은 서울 지역의 일별 오존 농도로 서울특별시의 대기 오염 물질과 기상 요소를 분석 자료로 사용하였다. 오존을 포함한 대기 오염 물질 자료(CO, NO_2, PM_{10}, SO_2)는 2005년 1월 1일부터 2007년 12월 31일까지의 3년간의 자료로 서울특별시 보건 환경 연구원에서 얻은 자료이다. 서울특별시에서는 27개의 대기 오염 측정소에서 측정이 이루어지는데 연구에서 사용한 자료는 27개 지점의 일 평균값이다. 기상요소는 기상청에 있는 여러 요소 중에서 오존과 관련이 있을 것으로 생각되는 기온, 강수량, 풍속, 습도, 일조시간에 대한 자료를 기상청으로부터 얻었다. 기상 요소 자료도 대기 오염 물질 자료와 마찬가지로 2005년 1월 1일부터 2007년 12월 31일까지의 3년간의 일평균 자료를 사용하였다.

2. 연구내용 및 방법

본 연구는 대기오염 물질 자료(CO, NO_2, PM_{10}, SO_2)와 기후인자 자료(기온, 강수량, 풍속, 습도, 일조시간)을 이용하여 서울지역 오존 농도 예측을 위한 모형을 개발하는 것이다. 오존 농도 예측모형을 개발하기에 앞서 오존농도의 연도별, 월별, 계절별, 시간별 변화 특성에 대해 알아보고 주요 대기오염 물질의 계절에 따른 시간별 특성과 오존농도와 기후인자, 대기오염물질 간의 상관관계에 대해 살펴보았다. 이를 토대로 다중회귀모형과 능형회귀모형을 이용한 오존농도 예측 모형을 개발하고 두 모형을 비교 검토하였다. 통계처리를 위한 패키지로는 SAS(Statisticla Analysis System) ver 9.1 , R ver 2.6.2을 사용하였다.

Ⅲ. 결과 및 고찰

1. 오존농도의 변화 특성

1.1 오존농도의 연도별 변화 특성

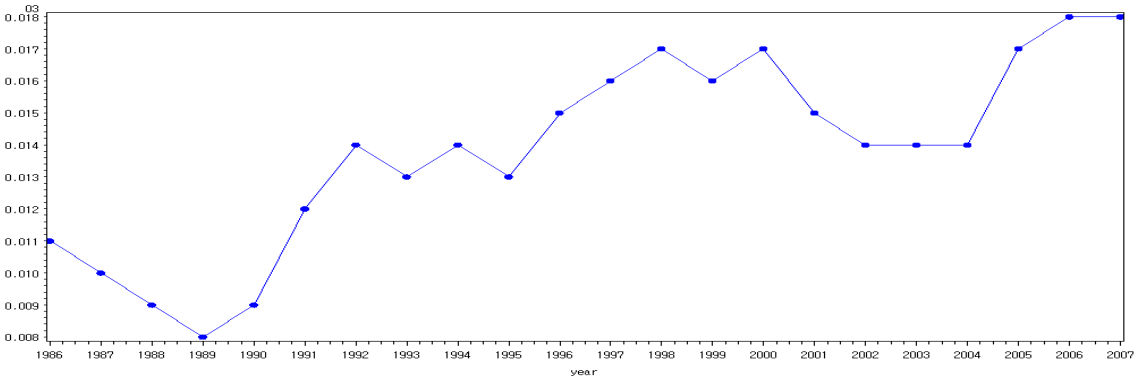

Fig. 1. O_3 농도의 연도별 변화 특성 (1986년~2007년)

서울지역의 1986년부터 2007년까지 오존농도의 연도별 변화 특성을 살펴보기 위해 연도별 오존농도의 평균값을 이용하여 Fig. 1에 나타내었다. 여기서 사용된 1986년부터 2006년까지의 오존 농도의 연평균 값은 국가통계포털 KOSIS에서 얻은 것이고, 2007년의 연평균 값은 분석에 사용된 자료를 이용하여 계산한 것이다. Fig. 1에서 살펴보면 서울지역 오존농도는 1986년부터 점차 감소하다 1989년 이후로 급격하게 증가하는 형태를 보이고, 2000년 이후로 감소하는 듯하나 2004년 이후로 다시 증가하는 형태를 나타낸다. 여전히 서울지역 오존농도는 높으며 오존농도 감소를 위한 대책 수립이 필요하다.

1.2 오존농도의 월별 변화 특성

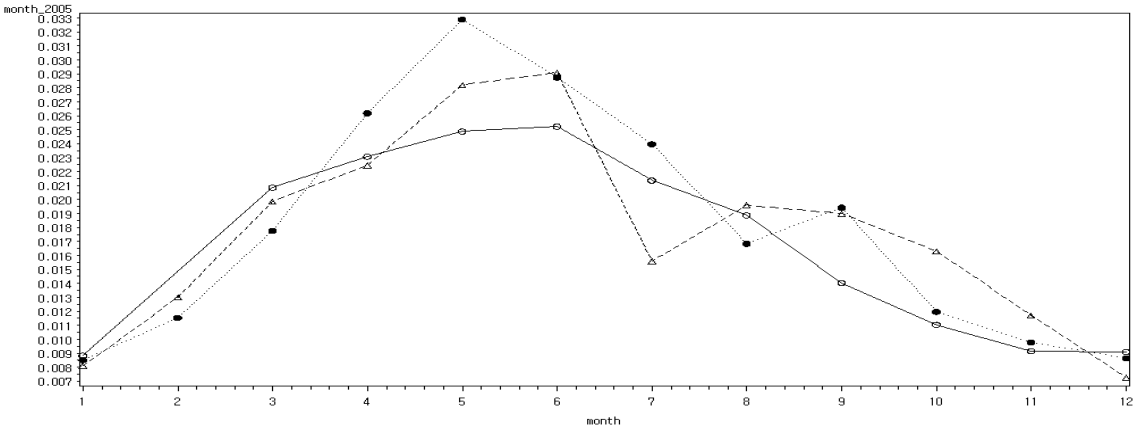

Fig. 2. O_3 농도의 월별 변화 특성(◦ : 2005년, △ : 2006년, ● : 2007년)

Fig. 2는 서울지역의 오존 농도 자료를 이용하여 월별 변화 특성을 파악한 것이다. 2005년과 2006년의 경우 6월에서 가장 오존 농도가 높으며, 2007년에는 5월에서 가장 농도가 높게 나타난다. 그 정점을 기준으로 오존 농도가 증가하다가 감소하는 형태를 띄는 것을 알 수 있다. 3년 평균 자료로 하였을 때는 5월의 농도가 가장 높은데 대전 지역 연구에 따르면 5월에는 습도가 낮고 일사량이 낮아 이와 같은 결과가 나타난다고 한다(정헌준 외, 2002). 5월의 오존 농도를 보면 해가 지날수록 농도가 높아지고 있는데 이는 연도별 특성을 보았을 때 2005년에서 2007년의 오존 농도가 점점 증가하는 것과 관련이 있다고 할 수 있겠다. 특이한 점은 2006년 7월의 오존 농도가 급격히 감소한 것인데 이는 2006년 7월의 총 강수량이 $1014mm$(기상청 홈페이지 내 과거자료 검색 참고)로 집중 호우가 영향을 주었을 것으로 사료된다. 전체적으로 보았을 때는 해가 지날수록 5월의 오존 농도가 급격히 증가하는 것을 알 수 있다.

1.3 오존농도의 계절변화 특성

봄을 3,4,5월, 여름을 6,7,8월, 가을을 9,10,11월, 겨울을 12,1,2월로 보고 각 계절별 오존 농도의 특성을 살펴보았다. Fig. 3~5는 순서대로 2005년부터 2007년까지의 계절별 오존 농도를 상자 그림으로 나타낸 것이다. 3년 모두 봄철의 오존 농도가 가장 높고, 겨울의 오존 농도가 가장 낮은 것을 알 수 있다. 봄철의 오존 농도가 높은 것은 습도가 낮고 일사량이 많은 5월이 포함되어 있어 계절 평균값이 높아진 것으로 생각된다. 연도별로 비교하면

2006년의 여름과 가을이 다른 연도에 비해서 오존 농도의 범위가 넓으며, 봄의 최대값은 2007년이 가장 크다.

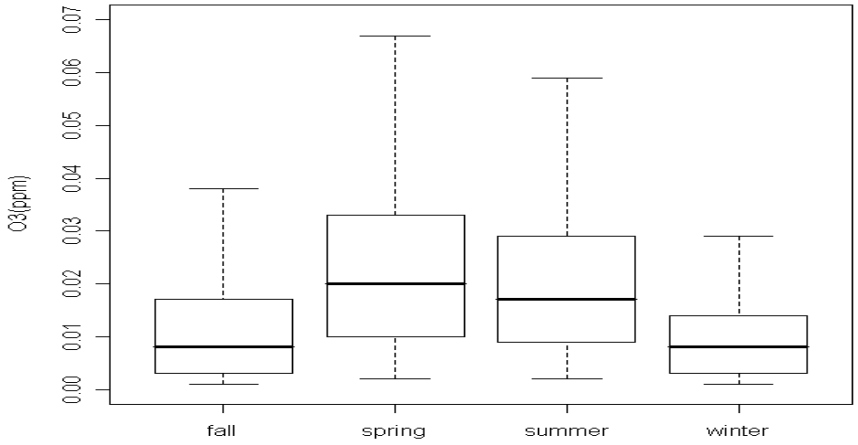

Fig 3. O_3 농도의 계절변화 특성(2005년)

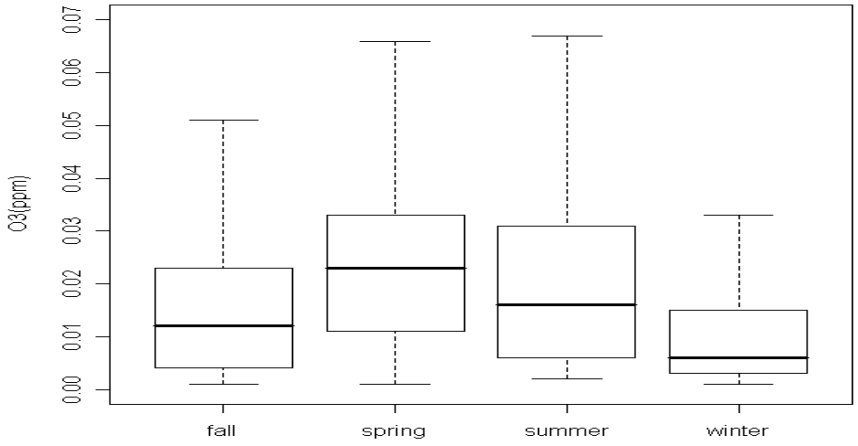

Fig 4. O_3 농도의 계절변화 특성(2006년)

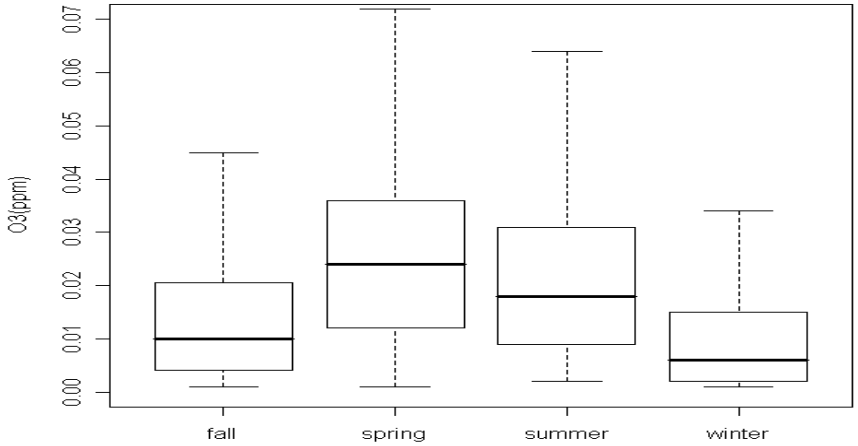

Fig. 5. O_3 농도의 계절변화 특성(2007년)

1.4 오존농도의 시간변화 특성

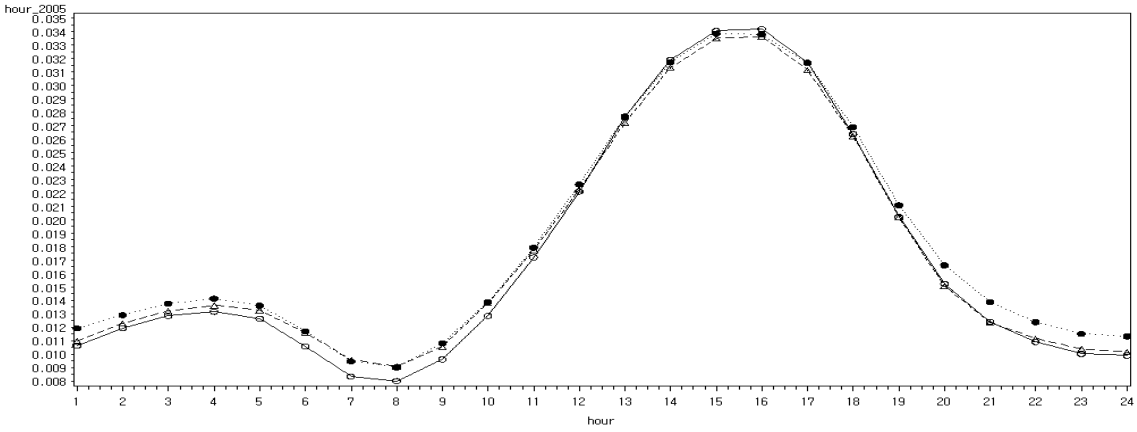

Fig. 6. O_3 농도의 시간변화 특성(◦ : 2005년, △ : 2006년, ● : 2007년)

오존농도의 시간별 평균 자료를 이용하여 Fig. 6에서 시간변화 특성에 대해 살펴보면 일일 중 오존농도는 오전 7시에서 9시 사이에 가장 낮게 나타나며 오전 9시 이후로 점차 증가하다가 오후 3시에서 4시 사이에 최대를 나타낸다. 이러한 현상은 오전에 교통량이 증가함에 따라 배출된 NO_2가 태양광선인 일사량의 영향으로 오전에 NO_2의 농도가 높아졌다가 3~4시간 후의 오존농도에 영향을 미치는 것이다 (김광진, 1998). 3년 동안의 시간별 오존 농도는 크게 차이 나지는 않으나 새벽 시간대나 저녁 시간대에서 2007년이 다른 연도에 비해 오존농도가 높게 나타나는 것을 알 수 있다.

2. 주요 대기오염물질의 시간별, 계절별 특징

오존의 농도 변화특성에서 월별로 농도에 차이가 나타나는 것을 바탕으로 오존을 포함한 대기 오염 물질인 CO, NO_2, PM_{10}, SO_2의 시간별 농도 변화를 계절에 따라 분석해 보았다. 여기서 쓰인 데이터는 2005년부터 2007년까지의 3년간의 대기오염 물질 자료이며 봄은 3,4,5월, 여름은 6,7,8월, 가을은 9,10,11월, 겨울은 12,1,2월을 나타낸다.

Fig. 7의 O_3 농도 변화를 보면 오후 3시와 4시 사이에 최대 농도를 보이며, 오후 2시부터 6시 사이를 제외하고는 봄철의 오존 농도가 가장 높고, 겨울철에 가장 낮다. 오후 1시부터 6시 사이의 오존농도가 여름에 가장 높고 모든 계절에서 해가 떠 있는 시간에 오존의 농도가 높은 것으로 보아 오존의 농도는 일조시간과 관계가 깊은 것으로 파악된다.

Fig. 8의 NO_2 농도 변화를 보면 오전 8시부터 10시 사이에 농도가 높게 나타나며 점차 감

소하는 경향을 보이다가 오후 4시 이후로 다시 증가하는 경향을 보인다. 계절별로는 겨울, 봄, 가을, 여름의 순으로 농도가 높았다. 여름의 경우에는 다른 계절과 다르게 오후 시간대보다 새벽 시간대의 농도가 확연히 낮다.

Fig. 9을 보면 SO_2 농도는 겨울철에 가장 높으며 시간별 변화는 겨울철에만 오후 12시 전후로 높은 농도를 보이고 있다. 봄, 여름, 가을에는 시간별 변화가 뚜렷하게 나타나지 않지만 낮에 비해 새벽의 농도가 낮게 나타난다.

Fig. 10의 CO 농도 변화를 보면 오전 9시에서 10시 사이에 최고농도를 나타내며 서서히 감소하다 오후 4시부터 다시 증가하는 경향을 보인다. CO 농도 역시 NO_2, SO_2와 마찬가지로 겨울철이 가장 높으며 여름철이 가장 낮다. 봄과 가을의 CO 농도는 거의 비슷하게 나타나고 있다.

마지막으로 Fig. 11에서 PM_{10}의 농도 변화 경향을 살펴보면 다른 대기 오염 물질과 다른 변화 형태를 가지는 것을 알 수 있다. 봄철에 가장 농도가 높으며, 여름철과 가을철에 농도가 가장 낮음을 알 수 있다. 봄철에 가장 농도가 높게 나타나는 것은 황사의 영향으로 보이며 시간별 변화는 특별한 형태가 나타나지 않는다.

Fig. 7~Fig. 11를 종합해 보면 오존 농도는 봄, 여름에 가장 높고 겨울에 가장 낮은 것으로 나타나나 PM_{10}을 제외한 나머지 대기오염물질은 오존 농도와 반대로 겨울에 농도가 가장 높고 여름에 농도가 가장 낮은 것으로 나타난다. 시간별 변화도 마찬가지로 오존과 대조적인 양상을 나타냄을 알 수 있다. 이는 오전 중에 오존전조물질의 축적이 있을 때 오존 농도가 높게 나타난다는 보고와 일치한다. (Angle, 1989)

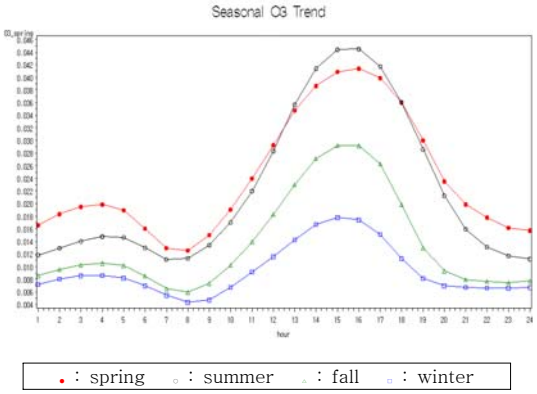

Fig. 7. 계절별 O_3농도의 시간별 변화

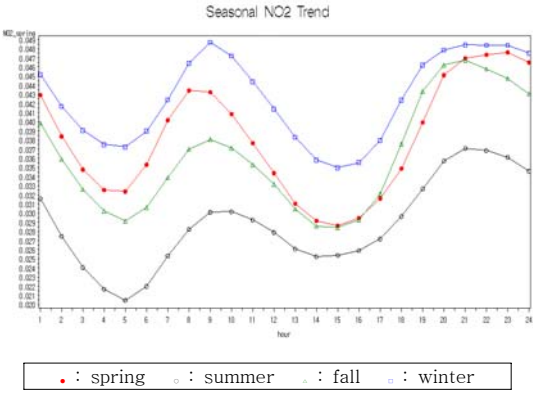

Fig. 8. 계절별 NO_2농도의 시간별 변화

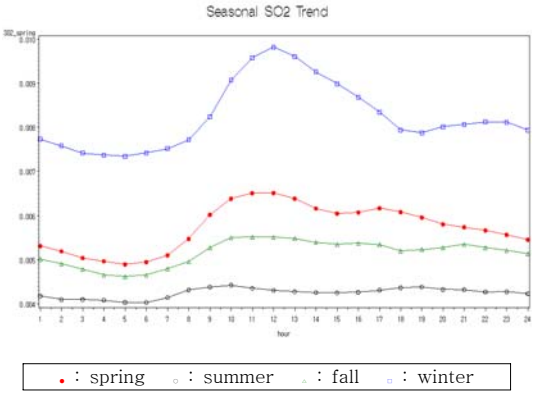

Fig. 9. 계절별 SO_2농도의 시간별 변화

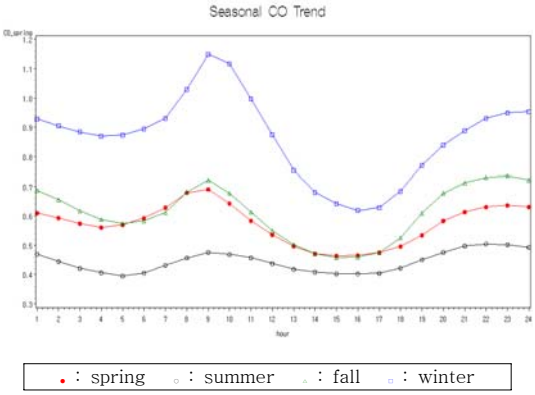

Fig. 10. 계절별 CO농도의 시간별 변화

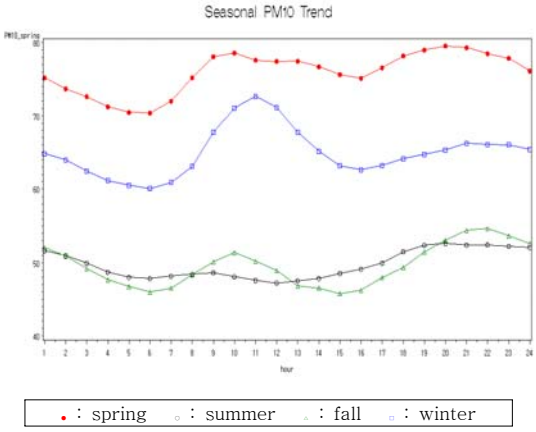

Fig. 11. 계절별 PM_{10}농도의 시간별 변화

3. 오존농도와 대기오염물질, 기후인자 간의 상관관계

서울지역의 오존농도와 대기오염물질 및 기후인자간의 관련성을 분석하기 위해서 상관분석을 실시하여 그 결과를 Table 1에 나타내었다. Table 1에서 오존농도와 대기오염물질 및 기후인자간의 상관관계를 살펴보면 오존농도는 대기오염물질 중 SO_2, CO, NO_2 와 음의 상관관계를 가지는 것으로 나타났으며, 그 상관계수는 각각 -.108, -.417, -.419이다. 그 중 NO_2와 CO는 상관계수가 -.419, -.417로 오존과 음의 상관성이 높으며 오존생성에 가장 큰 영향을 미침을 알 수 있다. 오존농도와 기후인자간의 상관관계를 살펴보면 풍속과 가장 높은 양의 상관관계를 가지고 있으며 기온과 일조시간 또한 상관계수가 각각 .382, .269로써 양의 상관성이 높으며 습도와도 상관계수 -.332로 비교적 높은 음의 상관성이 나타난다. 또한 상관분석 결과는 대기오염물질 간에 상관성이 높음을 나타내고 있다.

그러나 PM_{10}, 강수량과 오존농도 간에는 상관계수가 각각 .036, -.046으로 거의 상관관계가 없는 것으로 나타났으며 PM_{10}은 다른 대기오염물질과는 상관성이 있는 것으로 보이나 기후인자 간에는 상관성이 없는 것으로 나타났다. 강수량의 경우 직관적으로 오존에 많은 영향을 미칠 것으로 예상되나 상관계수는 거의 상관성이 없는 것으로 나타나는 것은 계절적 특성을 고려하지 않았기 때문으로 보인다.

Table 1. 오존농도와 대기오염물질, 기후인자 간의 상관계수

	O_3	PM_{10}	SO_2	CO	NO_2	강수량	일조시간	기온	풍속	습도
O_3	1									
PM_{10}	.036	1								
SO_2	-.108	.383	1							
CO	-.417	.370	.649	1						
NO_2	-.419	.315	.556	.782	1					
강수량	-.046	-.100	-.129	-.062	-.062	1				
일조시간	.269	-.009	.159	-.028	-.101	-.067	1			
기온	.382	-.079	-.341	-.401	-.240	.071	-.112	1		
풍속	.414	-.036	-.074	-.356	-.409	-.001	.143	-.001	1	
습도	-.332	-.037	-.179	.063	-.042	.279	-.633	.225	-.209	1

앞서 오존 농도의 계절변화 특성에서 살펴본 바와 같이 오존 농도는 계절별로 큰 차이를 나타내기 때문에 Table 2~5에서는 계절별로 오존농도와 대기오염물질, 기후인자 간의 상관관계를 살펴보았다.

Table 2는 봄의 상관계수 표이다. 계절별로 나누기 전과 비교하면 SO_2가 양의 상관을 보

이는데 상관계수가 .054로 거의 관계가 없는 것으로 나타났으며 반면 습도의 경우에는 상관계수가 -.332로 계절별로 나누기 전보다 음의 상관관계가 높아진 것을 알 수 있다.

Table 3은 여름의 상관계수 표이다. 오존과 다른 변수들과의 상관 계수를 보면 계절별로 나누기 전과 비교하여 PM_{10}과 강수량의 상관정도가 높아졌으며 오존과 음의 상관관계를 가졌던 SO_2, CO, NO_2는 양의 상관을, 양의 상관관계를 가졌던 기온과 풍속은 음의 상관을 나타내고 있다.

Table 4는 가을의 상관계수 표이다. 계절별로 나누기 전과 비교하여 PM_{10}이 음의 상관으로 나타났으며, SO_2의 상관 정도가 높아지고 반면 습도의 상관 정도는 낮아졌다.

Table 5는 겨울의 상관계수 표로 가을과 마찬가지로 PM_{10}이 오존과 음의 상관을 보였으며 CO, NO_2, 풍속과 오존의 상관 정도가 다른 계절에 비해 매우 크게 나타나고 있다. 기온의 경우 계절별로 나누기 전에는 양의 상관관계를 가졌으나 겨울에는 음의 상관관계를 가지고 있다.

계절별로 나누어서 상관 분석을 한 결과 대기 오염 물질들 간의 상관 계수가 높아졌으며, 각 계절별로 다른 상관관계를 가지는 것을 알 수 있다.

Table 2. 봄철 오존농도와 대기오염물질, 기후인자 간의 상관계수

	O_3	PM_{10}	SO_2	CO	NO_2	강수량	일조시간	기온	풍속	습도
O_3	1									
PM_{10}	.028	1								
SO_2	.054	.180	1							
CO	-.324	.263	.637	1						
NO_2	-.483	.070	.574	.714	1					
강수량	-.185	-.210	-.299	-.087	-.085	1				
일조시간	.357	-.062	.098	-.127	-.093	-.360	1			
기온	.354	.039	-.007	-.032	.190	.242	.043	1		
풍속	.358	-.004	-.200	-.457	-.677	.103	-.063	-.334	1	
습도	-.076	-.050	-.048	.122	-.088	.443	-.591	.079	.131	1

Table 3. 여름철 오존농도와 대기오염물질, 기후인자 간의 상관계수

	O_3	PM_{10}	SO_2	CO	NO_2	강수량	일조시간	기온	풍속	습도
O_3	1									
PM_{10}	.458	1								
SO_2	.335	.643	1							
CO	.324	.681	.629	1						
NO_2	.319	.532	.661	.809	1					
강수량	-.311	-.197	-.233	.022	-.012	1				
일조시간	.493	.052	.270	.051	.185	-.274	1			
기온	-.039	.101	.186	-.036	-.137	-.235	.287	1		
풍속	-.107	-.305	-.343	.531	-.561	.039	-.159	-.157	1	
습도	-.550	-.091	-.338	-.066	-.295	.466	-.706	-.180	.083	1

Table 4. 가을철 오존농도와 대기오염물질, 기후인자 간의 상관계수

	O_3	PM_{10}	SO_2	CO	NO_2	강수량	일조시간	기온	풍속	습도
O_3	1									
PM_{10}	-.149	1								
SO_2	-.354	.754	1							
CO	-.468	.784	.859	1						
NO_2	-.514	.673	.798	.900	1					
강수량	-.011	-.162	-.227	-.166	-.187	1				
일조시간	.179	-.052	.014	-.006	.044	-.318	1			
기온	.480	-.159	-.378	-.325	-.268	.152	-.157	1		
풍속	.400	-.269	-.248	-.464	-.531	.203	-.126	-.051	1	
습도	-.036	.114	-.032	.081	.014	.474	-.640	.429	-.056	1

Table 5. 겨울철 오존농도와 대기오염물질, 기후인자 간의 상관계수

	O_3	PM_{10}	SO_2	CO	NO_2	강수량	일조시간	기온	풍속	습도
O_3	1									
PM_{10}	-.248	1								
SO_2	-.469	.667	1							
CO	-.715	.599		1						
NO_2	-.795	.515	.692	.896	1					
강수량	-.089	-.134	-.150	-.010	-.014	1				
일조시간	.434	-.292	-.306	-.304	-.309	-.330	1			
기온	-.380	.410	.463	.574	.602	.144	-.445	1		
풍속	.743	-.174	-.416	-.667	-.730	.096	.107	-.395	1	
습도	-.345	.350	.338	.337	.296	.422	-.622	.460	-.107	1

4. 오존농도 예측모델의 개발

4.1 다중회귀모형

오존농도에 대한 예측 모형을 구축하기 위해 기상 요소와 대기 오염 요소를 가지고 다중회귀 분석을 하였다. 이 때, 오존 농도의 월별, 계절별 특성을 감안하여 계절에 따라 모형을 구축하였다. 계절은 봄을 3월에서 5월, 여름을 6월에서 8월, 가을을 9월에서 11월, 겨울은 12월에서 2월로 설정하였다. 모형구축에 사용한 변수는 9개의 대기오염물질과 기상인자들 중 각 계절별 상관분석 결과로부터 오존 농도와 상관관계가 없는 변수들은 제외한 후 단계적 선택법(stepwise)을 사용하였다. 각 모형에 사용된 독립변수는 Table 6과 같다.

Table 6. 계절별 모형구축에 사용된 독립변수

	독립변수
봄	CO, NO_2, 기온, 풍속, 일조시간
여름	PM_{10}, SO_2, CO, NO_2, 습도, 일조시간, 강수량
가을	PM_{10}, SO_2, CO, NO_2, 기온, 풍속, 일조시간
겨울	PM_{10}, SO_2, CO, NO_2, 기온, 습도, 풍속, 일조시간

위의 독립변수를 이용한 계절별 오존 농도의 다중 회귀식은 다음과 같다.

봄 : $O_3 = 0.00716 + 0.01537 \times CO - 0.42844 \times NO_2 + 0.00081 \times temp + 0.00370 \times Wind + 0.00074 \times Sun$

여름 : $O_3 = 0.04588 + 0.00016 \times PM_{10} - 2.61203 \times SO_2 - 0.00033 \times Humidity + 0.00103 \times Sun$

가을 : $O_3 = 0.00196 + 0.00008 \times PM_{10} - 0.25099 \times NO_2 + 0.00049 \times Temp + 0.00291 \times Wind + 0.00062 \times Sun$

겨울 : $O_3 = 0.01281 + 0.00003 PM_{10} - 0.25651 \times NO_2 + 0.00030 \times Temp - 0.00004 \times Humidity + 0.00226 \times Wind + 0.00046 \times Sun$

각 회귀식의 수정된 R^2값은 봄은 0.5966, 여름은 0.5153, 가을은 0.5944, 겨울은 0.7936로써 여름을 제외한 나머지 모든 계절에서 NO_2가 오존농도에 가장 큰 기여를 하는 것으로 나타났다. 다중회귀모형을 이용하여 얻어진 예측값과 실제 오존 농도간의 상관관계를 Fig. 12에 나타내었다. 그래프에서 X축은 예측값이고, Y축은 실제값을 나타내는데 선형 형태일수록 예측이 잘 되었다고 할 수 있다. 네 계절을 비교해 보면 겨울이 예측이 가장 잘 되었고, 여름은 그에 비해 예측력이 떨어진다.

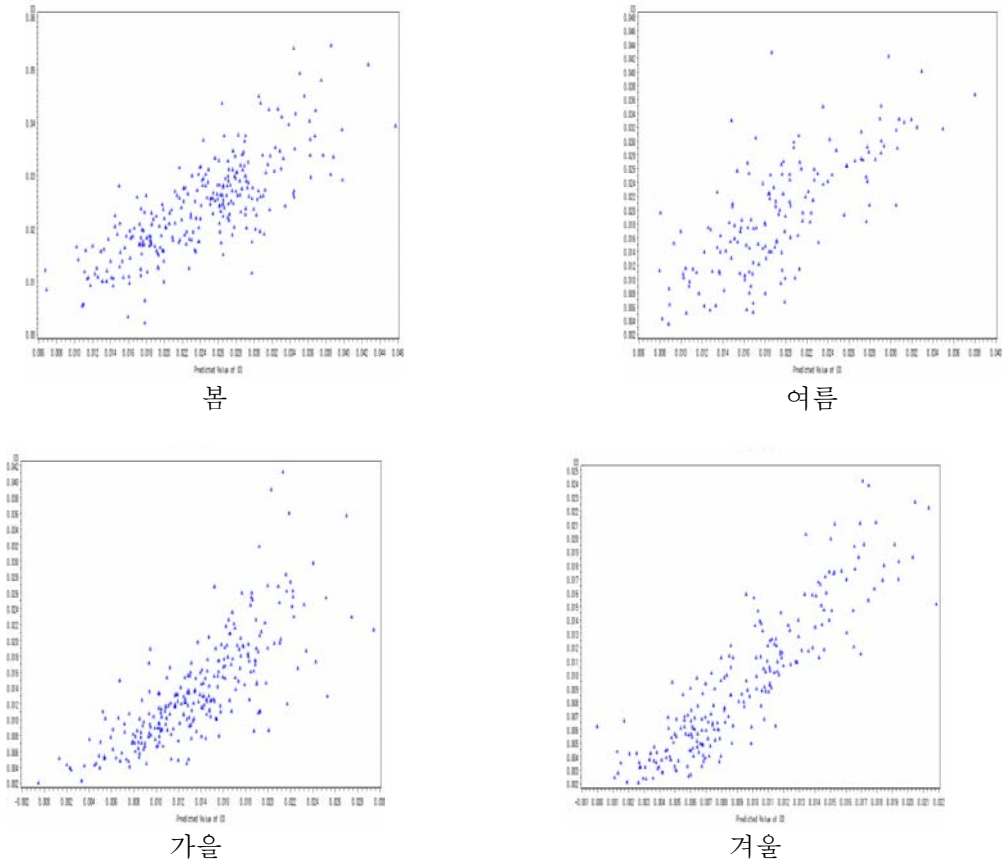

Fig. 12. 다중회귀를 이용한 예측치와 실제 오존 농도간의 관계

4.2 능형회귀모형

앞서 오존농도와 대기오염물질, 기후인자 간의 상관관계에 대해 살펴본 결과, 대기오염물질 간에도 강한 상관성이 있음을 확인할 수 있었다. 이는 다중회귀모형에서 다중공선성의 원인으로써 변수 선택법을 이용하거나 변수변환 등을 이용하여 문제를 해결할 수 있으나 본 연구에서는 다중공선성 문제 해결을 위한 또 다른 방안인 능형회귀모형을 구축하여 오존 농도를 예측해 보았다. 능형회귀모형 또한 다중회귀모형에서와 같이 오존농도의 계절별 특성을 고려하여 봄, 여름, 가을, 겨울 네 계절로 나누어 모형을 구축하였으며 모형 구축에 사용한 변수 또한 Table 6에서 다중회귀모형의 변수와 동일하다.

계절별 오존 농도의 능형 회귀식은 다음과 같다.

봄 : $O_3 = 0.01239 + 0.00577 \times CO - 0.30207 \times NO_2 + 0.00063 \times Temp + 0.00305 \times Wind + 0.00063 \times Sun$

여름 : $O_3 = 0.03771 + 0.00012 \times PM_{10} - 1.51286 \times SO_2 + 0.00138 \times CO + 0.03073 \times NO_2 - 0.00026 \times Humidity + 0.00092 \times Sun - 0.00001 \times Rainfall$

가을 : $O_3 = 0.00493 + 0.00005 \times PM_{10} + 0.04452 \times SO_2 - 0.00402 \times CO - 0.14789 \times NO_2 0.00039 \times Temp + 0.00240 \times Wind + 0.00046 \times Sun$

겨울 : $O_3 = 0.01116 + 0.00002 \times PM_{10} 0.04623 \times SO_2 - 0.00312 \times CO - 0.14475 \times NO_2 + 0.00016 \times Temp - 0.00003 \times Humidity + 0.00221 \times Wind + 0.00037 \times Sun$

4.3 모형비교

Fig 13~Fig. 16은 중회귀 모형에서 얻은 예측값과 능형회귀모형에서 얻은 예측값을 실제 오존농도와 비교한 그래프이다. 실선은 실제 오존의 농도를 나타내며 점선은 예측된 오존 농도를 의미한다. 실선과 점선이 거의 비슷할수록 예측이 잘 되었다고 할 수 있는데 부분적으로 직선이 나타나는 것은 강수량이 결측인 경우 예측이 되지 않아 결측값이 생성되어 나타난 결과이다. 또한 이 결과에서 살펴보면 오존의 농도가 고농도일 때 예측오차가 크게 발생되고 있다. 특히 오존의 농도가 높은 여름이나 가을의 예측값에서 모형 적합이 잘 되지 않음을 살펴볼 수 있으나 두 모형 모두 비교적 양호한 편이라 할 수 있을 것이다.

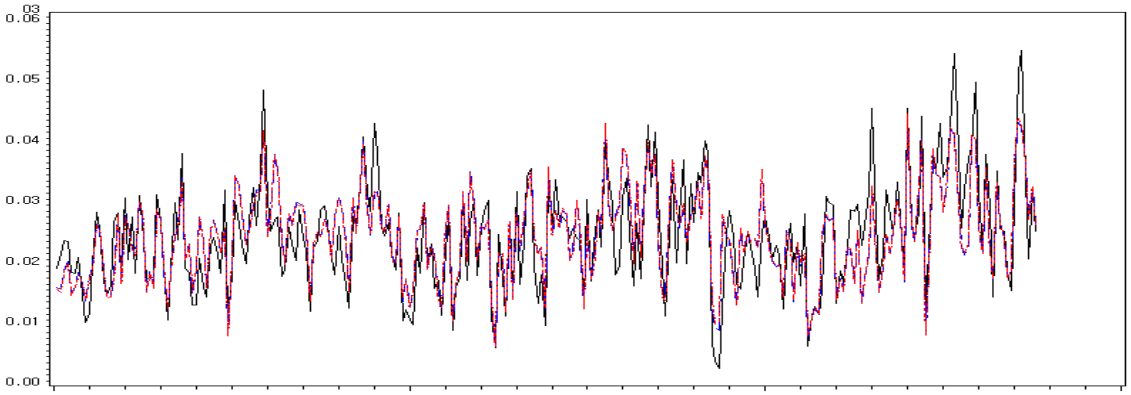

Fig. 13. 실제 오존 농도와 예측값의 비교 (봄)

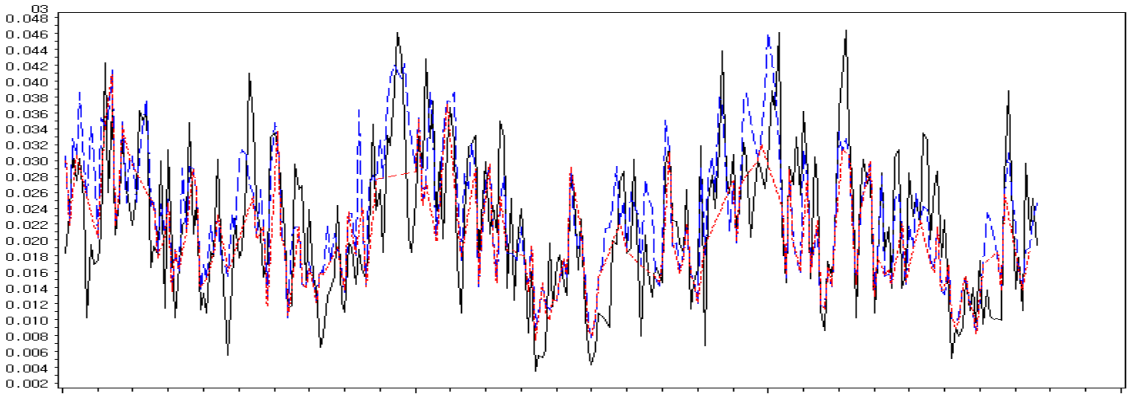

Fig. 14. 실제 오존 농도와 예측값의 비교 (여름)

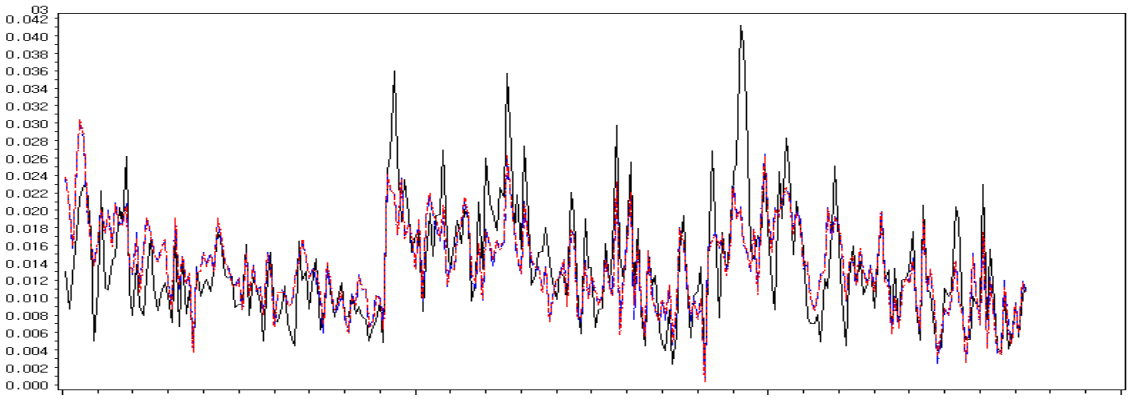

Fig. 15. 실제 오존 농도와 예측값의 비교 (가을)

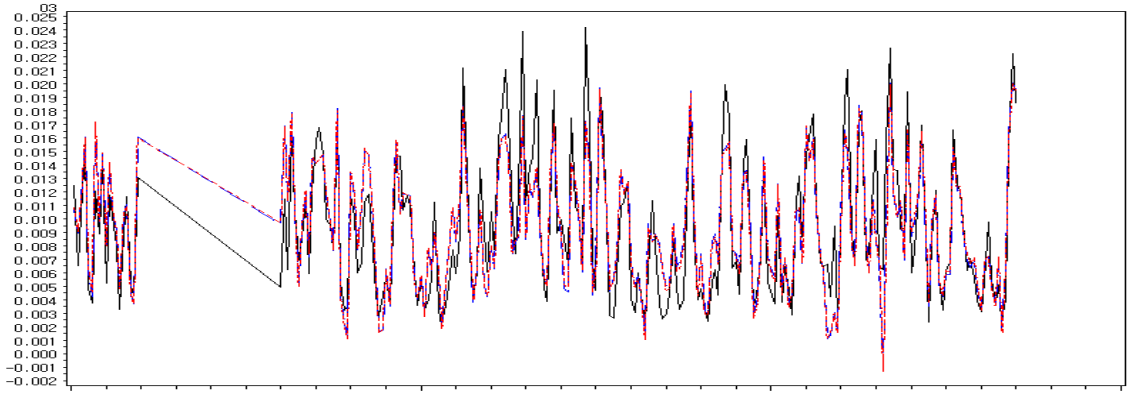

Fig. 16. 실제 오존 농도와 예측값의 비교 (겨울)

Ⅳ. 결론 및 향후과제

본 연구에서는 지구온난화 방지를 위한 하나의 대책으로 최근 매년 증가하고 있는 오존 농도를 예측하기 위해서 서울지역 27개 측정소의 대기오염 자료(CO, NO_2, PM_{10}, SO_2)와 기후인자 자료(기온, 강수량, 풍속, 습도, 일조시간)를 이용하여 예측 모형을 구축하였다. 자료는 모두 일별 평균값을 이용하였으며 오존농도의 연도별, 월별, 계절별, 시간별 변화특성과 주요 대기오염 물질의 계절에 따른 시간별 특성을 알아보았다. 오존의 농도는 계절별로 차이가 나타나는 것을 고려하여 오존 농도와 기후 인자, 대기 오염 물질 간의 상관 분석을 전체와 계절별로 실시한 후, 각각의 계절에서 오존과 상관이 있는 변수들을 일차적으로 선택하고 이 변수들을 토대로 다중 회귀 모형과 능형 회귀 모형을 이용한 오존 예측 모형을 개발하여 다음과 같은 결론을 얻었다.

1) 서울지역 오존 농도는 1986년부터 점차 감소하다 1989년 이후 급격하게 증가하여 2000년 이후 다시 감소하는 듯하다 2004년 이후로 2007년까지 점점 증가하는 추세이며, 2005년부터 2007년까지의 3년간의 월평균 오존 농도를 비교해 보면 5월의 오존 농도가 가장 높고, 12월의 오존 농도가 가장 낮다. 5월에 오존 농도가 높은 것은 습도가 낮고 일사량이 많은 5월의 기후적 특성이 영향을 준 것으로 생각된다. 오존농도의 계절변화 특성은 3월~5월을 봄, 6월~8월을 여름, 9월~11월을 가을, 12월~2월을 겨울로 보았을 때 봄철의 오존농도가 가장 높으며 겨울철의 오존농도가 가장 낮은 것으로 나타났다. 이는 5월 오존농도의 영향으로 계절 평균이 높아진 것으로 보이며 이는 온도와 일사량이 높고 습도가 낮을 때 오존 농도가 높아짐을 시사한다. 오존의 시간별 평균 자료를 가지고 분석한 결과 일일 중 오존농도는 오전 7시에서 9시 사이에 가장 낮게 나타나며 오전 9시 이후로 점차 증가하다가 오후 3시에서 4시 사이에 최대를 나타내고 일사량이 거의 없는 오후 8시 이후 감소하는 경향을 나타낸다.

2) 오존을 포함한 대기 오염 물질인 CO, NO_2, PM_{10}, SO_2의 시간별 농도 변화를 계절에 따라 분석해 보았다. 오존 농도는 오후 3시와 4시 사이에 최대 농도를 보이며 오후 1시에서 6시 사이에는 여름철의 오존 농도가 가장 높고 그 외의 시간대에서는 봄철의 오존 농도가 가장 높다. 나머지 계절은 가을, 겨울 순으로 오존 농도가 높다. 나머지 대기 오염 물질 중 CO, NO_2, SO_2은 오존 농도와 반대로 겨울에 농도가 가장 높고 여름에 농도가 가장 낮은 대조적인 양상을 나타낸다.

3) 오존 농도와 대기오염물질, 기후인자 간의 상관관계 결과 NO_2와 CO는 상관계수가 각

각 -.419, -.417로 강한 음의 상관을 가지는 반면 PM_{10}과 강수량은 오존농도와 상관계수가 각각 .036, -.046으로 거의 상관관계가 없는 것으로 나타난다. 오존농도의 계절적 특성을 고려하여 계절별로 오존 농도와 대기오염물질, 기후인자 간의 상관관계에 대해 분석해보면 봄의 오존농도는 PM_{10}, SO_2, 습도, 강수량과 관계가 없으며, 여름의 오존농도는 기온, 풍속과 상관관계가 없는 것으로 나타난다. 또한 가을의 오존농도는 습도, 강수량과 겨울의 오존농도는 강수량과 상관관계가 없다.

4) 오존 농도와 대기오염물질, 기후인자 간의 상관관계를 고려한 변수 선택한 후 다중회귀모형과 능형회귀모형을 이용하여 오존 농도를 예측하였다.

계절별 오존 농도의 다중 회귀식은 다음과 같으며,

봄 : $O_3 = 0.00716 + 0.01537 \times CO - 0.42844 \times NO_2 + 0.00081 \times temp + 0.00370 \times Wind + 0.00074 \times Sun$

여름 : $O_3 = 0.04588 + 0.00016 \times PM_{10} - 2.61203 \times SO_2 - 0.00033 \times Humidity + 0.00103 \times Sun$

가을 : $O_3 = 0.00196 + 0.00008 \times PM_{10} - 0.25099 \times NO_2 + 0.00049 \times Temp + 0.00291 \times Wind + 0.00062 \times Sun$

겨울 : $O_3 = 0.01281 + 0.00003 PM_{10} - 0.25651 \times NO_2 + 0.00030 \times Temp - 0.00004 \times Humidity + 0.00226 \times Wind + 0.00046 \times Sun$

계절별 오존 농도의 능형 회귀식은 다음과 같다.

봄 : $O_3 = 0.01239 + 0.00577 \times CO - 0.30207 \times NO_2 + 0.00063 \times Temp + 0.00305 \times Wind + 0.00063 \times Sun$

여름 : $O_3 = 0.03771 + 0.00012 \times PM_{10} - 1.51286 \times SO_2 + 0.00138 \times CO + 0.03073 \times NO_2 - 0.00026 \times Humidity + 0.00092 \times Sun - 0.00001 \times Rainfall$

가을 : $O_3 = 0.00493 + 0.00005 \times PM_{10} + 0.04452 \times SO_2 - 0.00402 \times CO - 0.14789 \times NO_2 0.00039 \times Temp + 0.00240 \times Wind + 0.00046 \times Sun$

겨울 : $O_3 = 0.01116 + 0.00002 \times PM_{10} 0.04623 \times SO_2 - 0.00312 \times CO - 0.14475 \times NO_2 + 0.00016 \times Temp - 0.00003 \times Humidity + 0.00221 \times Wind + 0.00037 \times Sun$

위의 두 모형을 이용하여 오존 농도를 예측한 결과 비교적 예측력은 양호한 편이었으나 고농도 오존에 대해서는 예측력이 떨어짐을 알 수 있었다. 이는 오존 대상 권역을 나누지 않고 서울 전역을 한 지역으로 처리함으로써 오존 발생과 분포의 국지성을 무시한 것으로 향후 연구를 통해 살펴볼 것이다. 또한 본 연구에서는 일별 자료를 이용한 것으로 시간별 자료를 이용하여 오존 농도를 예측해 보는 것도 의미가 있을 것이다.

Reference

[1] Angle, R. P. and H. S. Sandhu (1989) Urban and rural ozone concentrations in Alberta, Canada, Atmos. Environ., 23(1), 215-221.

[2] 김광진. (1998). 수도권 지역의 광화학 오존농도 저감방안에 관한 연구, 건국대학교 대학원 석사학위논문.

[3] 서의훈, 정연선. (2001). 오존농도의 예측 모형에 관한 연구, The Journal of Korean Data Analysis Society, 3(3), 319-330.

[4] 윤오섭, (1999). 최신환경학, 세진사, 171-173.

[5] 전병일, 김유근, 이화운. (1995). 부산지역의 오존 농도 특성과 기상 인자에 관한 연구, 한국대기보전학회지, 11(1), 45-46.

[6] 정헌준, 백승화, 김종헌. (2002). 대전지역 대기 중 오존예보에 관한 연구, 환경관리학회지, 8(2), 131-144.

[7] 최성우, 최상기, 도상현. (2002). 다중회귀분석을 통한 대구지역 오존농도 예측, 한국환경과학회지, 11(6), 687-696.

Annals of Statistical Research
Vol. 4. No. 1, 2008, pp.35~49

On meta-analysis for summarization of previous studies

Hongyup Ahn[1], Minsu Kang[2]

Abstract

For various scientific topics, many researches have been published in journals. Obviously, those previous research results are used for the current or later researches. However, there are discrepancy between results of previous studies for the same scientific topic. Whether such discrepancy is negligible or not, it is important that those various results should be combined as one result. Thus, the later studies could be consistently planned no matter which previous study is referred. For this, a summarization of many reported results has become a growing problem. Statistical approach has been considered to summarize previous works. It is meta-analysis. In this writing, we review several methodologies on meta-analysis and present an exemplary analysis.

Keywords : Forest plot, Funnel plot, Meta-analysis, Odds ratio, Random effects model, Relative risk.

1) (Corresponding author) Assistant Professor, Department of Statistics, Dongguk University, Seoul, 100-715, Korea. E-mail: ahn@dongguk.edu
2) Master's course, Department of Statistics, Dongguk University, Seoul 100-715, Korea

1. 서론

통계학은 다양한 학문 영역에서 자료 분석을 위해 광범위하게 사용되고 있다. 특정 주제에 관한 연구를 진행할 때 일반적으로 자료를 먼저 준비하고 이에 대한 적절한 통계모형을 선택하여 분석을 실시한다. 통계분석을 통해 얻어진 결론은 연구자의 자료에서 얻을 수 있는 결론의 일반화라고 할 수 있다. 하지만 동일한 연구 주제임에도 불구하고 실험설계의 차이 또는 구체적으로 관측된 자료의 차이로 인해 상이한 연구 결론이 발표되기도 한다. 이러한 연구 결과들은 발표로써 끝나는 것이 아니라 이들 연구 결과를 참고하는 다른 연구에 연향을 주게 된다. 선행연구들의 일관된 결과는 이어지는 다른 연구의 진행에 많은 도움을 주지만 일치되지 않는 연구 결과들은 정리의 필요성이 절실히 요구된다. 또한 일관된 결론이라도 이들 결과들에 대해 하나의 정량화된 결과가 요구되기도 한다.

Glass (1976)는 동일 연구 주제에 관한 개별 연구들의 결과를 통계적으로 분석하는 방법을 메타분석(Meta-analysis)이라 정의하였다. 개별 연구의 결과로써 발표되는 통계량은 평균, 상관계수, 유의확률, Odds Ratio(OR), Relative Risk(RR) 등 매우 다양하다. 이들 통계량과 더불어 신뢰구간, 표준편차 등과 같은 분포에 관한 결과들도 함께 제시되곤 한다. 메타분석에서는 통계량과 이에 대응되는 표준편차를 이용하여 먼저 그 결과들을 표준화시킨 후 이들의 대표값을 구하는 것에 그 목적을 두고 있다. Pearson (1904)은 상관계수에 대해, Tippet (1931)과 Fisher (1932)는 유의확률에 관한 메타분석을 각각 소개하였다.

메타분석을 할 때에는 메타분석에 이용되는 개별 연구들의 성격이 중요하다. 개별 연구의 결과들이 유사할 경우에는 이들을 하나의 대표값으로 나타내었을 때 그 값에 대해 신뢰를 할 수 있다. 반면, 유사하지 않은 개별 연구의 결과를 하나의 값으로 종합했을 경우에는 그 신뢰성이 저하되기 마련이다. 이 때문에 메타분석을 할 때 개별 연구들의 결과에 대해 동질성의 정도를 고려해야 한다. 만약 동질성이 만족되지 않는다면 개별 연구들의 차이를 반영한 메타분석을 해야 한다. 동질성의 정도를 검정하는 통계량으로 Hardy와 Thompson (1998)은 Q통계량을 소개했다. 또한, Higgins와 Thompson (2002)는 이질성을 판별할 수 있는 두 가지의 척도, H와 I^2를 소개하였다.

메타분석은 동질성의 만족 여부에 따라 두 개의 모형을 고려할 수 있다. 동질성이 만족되었다면 모수효과 모형을 사용하고, 동질성이 만족되지 못하였다면 랜덤효과 모형을 사용하여 메타분석을 한다 (DerSimonian and Laird, 1986).

메타분석의 분석 대상은 이미 발표된 논문의 연구 결과들이다. 이런 의미에서 연구 주제에 관련된 논문을 충분히 확보하는 것이 메타분석의 성공여부를 결정짓는다. 하지만 유의한

결과들이 발견된 논문이 주로 발표되는 경향이 있기 때문에 메타분석에 이용되는 논문들의 선택 과정에서 publication bias가 발생할 가능성이 있다 (Whitehead, 2002). 따라서 메타분석 결과의 신뢰성을 확보하는데 있어 publication bias의 존재유무를 확인하는 것은 반드시 필요한 단계라 할 수 있다.

본 논문에서는 메타분석에 관한 전반적인 내용과 실제 사용할 수 있는 프로그램을 소개하도록 하겠다. 2장에서 메타분석의 이론적인 부분을, 3장에서 메타분석을 수행하는 간단한 예를 준비하였다.

2. 메타분석

메타분석을 실시하기 위해 k개의 논문을 검색하였다고 하자. i번째 논문의 통계량과 통계량의 분산을 각각 $\hat{\theta}_i$와 $Var(\hat{\theta}_i) = w_i^{-1}$라 하자. 또한 k개 논문 결과들의 요약한 대표값 즉 효과크기를 $\hat{\theta}_{pooled}$라 하자.

평균. 평균을 대상으로 메타분석을 실시하는 경우, i번째 논문의 통계량 $\hat{\theta}_i$는 표본 평균 $\overline{x}_i$이고 w_i^{-1}는 표준오차 s_i^2/n_i로 추정이 된다. 여기서, s_i^2와 n_i는 각각 i번째 논문의 표본분산과 표본 크기다. $\hat{\theta}_{pooled}$은 다음과 같이 정의 된다.

$$\hat{\theta}_{pooled} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} n_i \hat{x}_i / s_i^2}{\sum_{i=1}^{k} n_i / s_i^2} \tag{1}$$

여기서 k개의 논문 결과를 하나의 값 $\hat{\theta}_{pooled}$으로 추정하는데 필요한 전제는 $\hat{\theta}_i$들이 유사한 값이라는 사실이다. 이에 대해 확인하기 위해 동질성 검정 ($H_0 : \theta_i = \theta_0, \quad i = 1,2,...,k$)을 다음과 같은 검정통계량을 통해 실시한다.

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}_{pooled}\right)^2 \tag{2}$$

식 (2)의 검정통계량 Q의 분포는 자유도 $k-1$인 χ^2분포로 근사된다. 따라서 $Q < \chi^2_{(\alpha,k-1)}$ 이면 유의수준 α하에 k개의 $\hat{\theta}_i$들은 동질하다고 볼 수 있고 이들의 대표값으로 $\hat{\theta}_{pooled}$을 사용할 수 있다.

만일 s_i^2와 n_i가 논문에서 제시되어 있지 않아 w_i를 계산할 수 없는 경우 신뢰구간 등을

이용하여 w_i를 계산할 수 있다. 예를 들면, 평균의 신뢰구간이 (a,b)로 주어졌다면 w_i는 다음과 같다.

$$w_i = \frac{2z_{\alpha/2}}{b-a} \tag{3}$$

메타분석의 결과로 $\hat{\theta}_{pooled}$와 함께 $\hat{\theta}_{pooled}$의 신뢰구간이 제시되어야 한다. 신뢰구간을 계산하기 위해 식 (1)에 대한 분산을 먼저 계산하면

$$Var(\hat{\theta}_{pooled}) = \left(\sum_{i=1}^{k} w_i\right)^{-1} \tag{4}$$

이 되고 $(1-\alpha)100\%$ 신뢰구간은 다음과 같이 구해진다.

$$\hat{\theta}_{pooled} \pm z_{\alpha/2}\left(\sum_{i=1}^{k} w_i\right)^{-1/2} \tag{5}$$

일반적으로 효과크기 $\hat{\theta}_i$가 0이 아닌 경우 효과의 존재를 의미한다. 이는 $\hat{\theta}_{pooled}$에 대해서도 마찬가지이다. 따라서 식 (5)의 신뢰구간이 0을 포함하지 않는다면 k개의 논문결과에 대한 메타분석 결과로 효과의 존재를 말할 수 있다.

Odds Ratio(OR). OR을 대상으로 메타분석을 실시할 경우는 발생빈도를 이용하여 비교하는 연구들의 결과를 이용한다. 예를 들면, 두 군에서의 암 발생과의 관계를 확인하기 위해 i번째 논문 결과를 [표 1]에서와 같은 2×2 분할표로 작성하여 OR을 계산하도록 한다.

[표 1] 2×2 분할표

	암 발병	정상
A 군	n_{11i}	n_{12i}
B 군	n_{21i}	n_{22i}

Reference level을 A 군이라 하고 암 발생에 대한 OR을 추정하면 다음과 같다.

$$\widehat{OR}_i = \frac{n_{21i}n_{12i}}{n_{11i}n_{22i}}$$

일반적으로 양수인 OR의 표본분표는 비대칭적이므로 정규근사를 이용한 추론을 위해서 로그변환을 이용한다. 또한 $\ln\widehat{OR}_i$의 asymptotic variance는

$$Var\left(\ln\widehat{OR}_i\right) \approx \frac{1}{n_{11i}} + \frac{1}{n_{12i}} + \frac{1}{n_{21i}} + \frac{1}{n_{22i}} \tag{6}$$

알려져 있다 (Agresti, 1996).

Mantel과 Haenszel (1959)은 $\hat{\theta}_i = \ln \widehat{OR}_i$들의 대표값으로 $\hat{\theta}_{pooled}$를 다음과 같이 제안하였다.

$$\ln \widehat{OR}_{MH} = \ln \left[\frac{\sum_{i=1}^{k} m_i \widehat{OR}_i}{\sum_{i=1}^{k} m_i} \right] = \ln \left[\frac{\sum_{i=1}^{k} \frac{n_{21i} n_{12i}}{n_i}}{\sum_{i=1}^{k} \frac{n_{11i} n_{22i}}{n_i}} \right] \tag{7}$$

여기서 $m_i = \frac{n_{11i} n_{22i}}{n_i}$이고 $n_i = n_{11i} + n_{12i} + n_{21i} + n_{22i}$이다. 평균에 대한 메타분석에서와 같이 식 (7)로 k개의 연구 결과를 대표하기 위해 동질성 검정이 필요하다. 동질성 검정은 식 (2)를 이용하여 실시하되 w_i, $\hat{\theta}_i$, $\hat{\theta}_{pooled}$를 각각 식 (6)의 $\left[Var\left(\ln \widehat{OR}_i\right) \right]^{-1}$, $\ln \widehat{OR}_i$, 식 (7)의 $\ln \widehat{OR}_{MH}$로 대체한다. 참고로 n_{11i}나 n_{22i}가 0인 경우 $\widehat{OR}_i$는 ∞가 되어 식 (7)의 $\widehat{OR}_i$을 이용한 공식을 사용하면 계산이 되지 않을 수 있다. 이에 반해 식 (7)의 마지막 식에서는 n_{11i}나 n_{22i}가 0이더라도 그 들의 합은 0이 되지 않으므로 문제되지 않는다.

만일 논문에서 2×2 분할표가 제시되지 않거나 작성을 위한 충분한 정보가 제공되지 않는 경우 식 (3)과 유사하게 $\widehat{OR}_i$(또는 $\ln \widehat{OR}_i$)의 신뢰구간을 이용하여 w_i를 계산할 수 있다. $\widehat{OR}_i$의 신뢰구간이 (a, b)로 주어졌다고 가정해 보면, $\ln \widehat{OR}_i$에 대한 신뢰구간은 $(\ln a,\ \ln b)$가 되므로 w_i는 다음과 같다.

$$w_i = \frac{2 z_{\alpha/2}}{\ln a + \ln b} \tag{8}$$

$\ln \widehat{OR}_{MH}$의 $(1-\alpha)100\%$ 신뢰구간을 구하기 위해 $\ln \widehat{OR}_{MH}$의 분산 추정량이 필요하다. Robins et al. (1986)은 다음과 같은 분산 추정량을 제안하였다.

$$Var(\ln \widehat{OR}_{MH}) = \frac{\sum_{i=1}^{k} P_i R_i}{2\left(\sum_{i=1}^{k} R_i\right)^2} + \frac{\sum_{i=1}^{k} (P_i S_i + Q_i R_i)}{2\left(\sum_{i=1}^{k} R_i\right) \cdot \left(\sum_{i=1}^{k} S_i\right)} + \frac{\sum_{i=1}^{k} Q_i S_i}{2\left(\sum_{i=1}^{k} S_i\right)^2} \tag{9}$$

여기서, $P_i = \frac{n_{12i} + n_{21i}}{n_i}$, $Q_i = \frac{n_{11i} + n_{22i}}{n_i}$, $R_i = \frac{n_{12i} n_{21i}}{n_i}$, $S_i = \frac{n_{11i} n_{22i}}{n_i}$이다. 그러므로 신뢰구간은

$$\ln \widehat{OR}_{MH} \pm z_{\alpha/2} \sqrt{Var\left(\ln \widehat{OR}_{MH}\right)}$$

이 되고 만일 신뢰구간이 0을 포함하지 않는다면 효과가 존재한다고 결론짓는다.

Relative Risk(RR). RR을 대상으로 메타분석을 실시할 경우, i번째 논문의 통계량 $\hat{\theta}_i$는 로그변환한 $\ln\widehat{RR}_i$를 사용한다. 여기서 [표 1]과 같은 분할표의 A 군을 reference level로 하여 암 발생에 대한 RR을 추정하면

$$\widehat{RR}_i = \frac{n_{21i}(n_{11i}+n_{12i})}{n_{11i}(n_{21i}+n_{22i})}$$

이 되고 $\ln\widehat{RR}_i$의 asymptotic variance는 $\frac{1}{n_{11i}} - \frac{1}{n_{11i}+n_{12i}} + \frac{1}{n_{21i}} - \frac{1}{n_{21i}+n_{22i}} \equiv w_i^{-1}$로 추정된다 (Fleiss, 1993).

Rothman (1982)은 $\hat{\theta}_i = \ln\widehat{RR}_i$들의 대표값으로 $\hat{\theta}_{pooled}$를 다음과 같이 제안하였다.

$$\ln\widehat{RR}_{MH} = \ln\left[\frac{\sum_{i=1}^{k}\frac{n_{21i}(n_{11i}+n_{12i})}{n_i}}{\sum_{i=1}^{k}\frac{n_{11i}(n_{21i}+n_{22i})}{n_i}}\right] \tag{10}$$

또한, Greenland와 Robins (1985)는 $\ln\widehat{RR}_{MH}$ 분산의 추정량을 다음과 같이 제시하였다.

$$Var(\ln\widehat{RR}_{MH}) = \frac{\sum_{i=1}^{k}\frac{(n_{11i}+n_{12i})(n_{21i}+n_{22i})(n_{11i}+n_{21i})}{n_i^2}}{\left(\sum_{i=1}^{k}\frac{n_{11i}(n_{21i}+n_{22i})}{n_i}\right)\cdot\left(\sum_{i=1}^{k}\frac{n_{21i}(n_{11i}+n_{12i})}{n_i}\right)} \tag{11}$$

동질성 검정은 OR에서와 유사하게 실시하고, $\ln\widehat{RR}_{MH}$의 신뢰구간은

$$\ln\widehat{RR}_{MH} \pm z_{\alpha/2}\sqrt{Var(\ln\widehat{RR}_{MH})}$$

이 되고 만일 신뢰구간이 0을 포함하지 않는다면 효과가 존재한다고 결론짓는다.

모형선택. 동질성 검정에서 귀무가설 $H_0 : \theta_i = \theta_0, \quad i = 1,2,...,k$이 채택되면 모수효과 모형을 선택하고, $\hat{\theta}_{pooled}$와 정규분포의 성질을 이용하여 $\hat{\theta}_{pooled}$의 $(1-\alpha)100\%$ 신뢰구간을 구할 수 있었다. 동질성 검정이 기각된 경우는 각 연구의 θ_i가 상이하다는 것이다. 이러한 경우에는 랜덤효과 모형을 사용하여 메타분석을 실시할 수 있다.

랜덤효과 모형에서 $\hat{\theta}_i$는 $N(\theta, w_i^{-1}+\tau^2)$를 가정한다. 이에 반해 모수효과 모형은 $\hat{\theta}_i$의 분포로 기대값과 분산이 각각 θ_i와 w_i^{-1}인 정규분포를 가정한다. 연구들 간의 차이를 각 모형에서 다르게 반영하고 있다. 모수효과 모형에서는 기대값의 차이로 랜덤효과 모형에서는 분산을 팽창시키는 요인의 형태로 설명되고 있다. 이러한 이유로 해서 모수효과 모형에서는

동질성 검정이 요구되고 동질성이 확보될 때에만 $\hat{\theta}_{pooled}$를 그 대표값으로 인정할 수 있다. 만일 동질성이 확보되지 않는다면 기대값이 동일하다는 가정 하에 연구 간의 차이를 기대값의 차이가 아닌 분포로 인해 발생한 것으로 보는 랜덤효과 모형을 이용하여 메타분석을 실시하게 된다.

랜덤효과 모형을 이용한 추정에서 연구 논문들 간의 효과차이를 τ^2를 통해 나타내고 그 추정량으로

$$\hat{\tau}^2 = (Q-k+1)/D \tag{12}$$

을 사용한다. 여기서 $D = \frac{k-1}{k\overline{w}}(k\overline{w}^2 - s_w^2) = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}$이고 $\overline{w}$와 s_w^2는 각각 w_i들의 평균과 분산이다 (Woodward, 2005). 따라서, $[Var(\hat{\theta}_i)]^{-1} = [w_i^{-1} + \tau^2]^{-1} \equiv w_i^{(R)}$이고 이를 가중치로 이용한 가중평균을 $\hat{\theta}_{pooled}$로 사용한다.

$$\hat{\theta}_{pooled} = \frac{\sum_{i=1}^{k} w_i^{(R)} \hat{\theta}_i}{\sum_{i=1}^{k} w_i^{(R)}} \tag{13}$$

또한 $(1-\alpha)100\%$ 신뢰구간도 식 (5)에서 w_i 대신에 $w_i^{(R)}$을 사용하여 다음과 같이 구해진다.

$$\hat{\theta}_{pooled} \pm z_{\alpha/2}\left(\sum_{i=1}^{k} w_i^{(R)}\right)^{-1/2} \tag{14}$$

Publication bias. 일반적으로 새로운 치료방법의 유의한 효과를 증명하는 연구결과가 아무런 효과가 없음을 증명한 연구결과보다 더 가치있게 받아들여진다. 이는 논문을 평가하고 발표하는 과정에 있어서도 많은 영향을 주어 통계적으로 유의한 결과를 보인 논문은 쉽게 인정되는 반면 유의하지 않은 결과의 논문은 발표되지 못할 가능성이 커지게 된다. 이러한 문제점으로 인해 메타분석에 사용되는 논문이 모든 연구결과들을 대표하지 못하게 되어 왜곡된 결과가 발생하게 된다. 그러므로 메타분석을 하는데 있어서 publication bias를 반드시 고려해야 한다.

Publication bias를 확인하는 방법으로는 funnel plot을 이용한다. Funnel plot의 x축은 각 논문 i의 효과크기 $\hat{\theta}_i$를, y축은 표본의 크기, 효과크기의 표준오차 또는 표준오차의 역

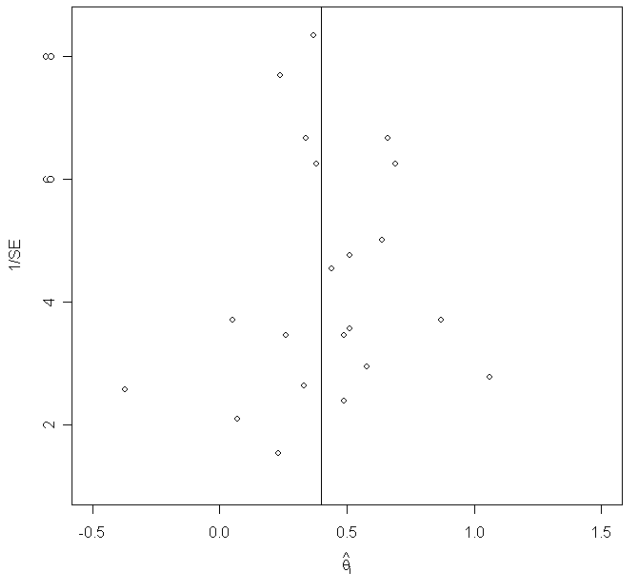

그림 1. Funnel plot.

수 등을 나타내어 그리는 산점도의 일종이다. Funnel plot의 일반적인 형태는 $\hat{\theta}_{pooled}$를 중심으로 좌우대칭인 삼각형의 모양을 보인다. 예를 들어 y축을 효과크기의 표준오차의 역수로 한다면 표본의 크기가 클수록 표준오차의 역수는 커질 것이다. 이와 동시에 $\hat{\theta}_i$는 $\hat{\theta}_{pooled}$에 더 가까이 분포할 것이다. 이러한 경향으로 인해 funnel plot의 형태는 [그림 1]과 같이 삼각형이 될 것이다. 하지만 funnel plot의 형태가 [그림 1]과 같은 삼각형의 형태를 띄지 않는다면 publication bias를 의심해 볼 수 있다.

Funnel plot과 같이 그래프를 이용하여 publication bias 존재 유무를 확인할 수 있지만 결과의 판단은 다소 주관적인 면이 있다. 이에 보조적인 방법으로 통계적 검정을 통해 확인하는 것을 생각할 수 있다. Begg와 Mazumdar (1994)는 순위상관방법을 Egger (1997)는 선형회귀방법을 publication bias를 확인하기 위해 고려했다.

[표 3] 34개의 콜레스테롤에 관한 RCT 결과 표.

논문 번호	실험군의 환자수	대조군의 환자수	실험군의 사망자수	대조군의 사망자수	논문 번호	실험군의 환자수	대조군의 환자수	실험군의 사망자수	대조군의 사망자수
1	204	202	28	51	18	350	367	42	48
2	285	147	70	38	19	79	78	4	5
3	156	119	37	40	20	1149	1129	37	48
4	88	30	2	3	21	221	237	39	28
5	30	33	0	3	22	54	26	8	1
6	279	276	61	82	23	71	72	5	7
7	206	206	41	55	24	4541	4516	269	248
8	123	129	20	24	25	421	417	49	62
9	1018	1015	111	113	26	94	94	0	1
10	427	143	81	27	27	311	317	19	12
11	244	253	31	51	28	1906	1900	68	71
12	50	50	17	12	29	2051	2030	44	43
13	47	48	23	20	30	6582	1663	33	3
14	30	60	0	4	31	5331	5296	236	181
15	5552	2789	1025	723	32	48	49	0	1
16	424	422	174	178	33	94	52	1	0
17	199	194	28	31	34	23	29	1	2

3. 메타분석 예제

앞 장에서 여러 통계량에 관한 메타분석을 살펴보았다. 가중 평균을 사용한 $\hat{\theta}_{pooled}$와 그에 따른 신뢰구간은 앞 장의 식을 이용하면 쉽게 구할 수 있으나 경우에 따라 계산의 실수가 발생할 여지가 있다. 그러므로 이러한 일련의 메타분석을 처리해 주는 통계프로그램을 이용하여 빠르고 정확하게 계산하는 것이 바람직하다. 메타분석에 사용되는 그림은 공개프로그램인 R의 rmeta라이브러리를 이용하여 작성하였다.

메타분석에 사용될 자료는 Smith et al. (1993)에서 제시한 콜레스테롤에 관련된 메타분석 자료이다. 34개의 randomized clinical trials (RCT) 결과는 [표 3]과 같이 실험군과 대조군의 환자 수와 각 군에서의 사망자 수를 정리하였다. 발생빈도의 비교를 통해 실험군과 대조군의 차이를 알아보는 것이므로 OR을 θ_i로 놓고 메타분석을 실시한다.

[표 4] 34개의 콜레스테롤에 관한 RCT의 $\widehat{OR}_i$, w_i, m_i, $w_i^{(R)}$ 표.

논문번호	$\widehat{OR}_i$	w_i	m_i	$w_i^{(R)}$
1	0.4710339	14.7871107	22.1083744	7.1521081
2	0.9339045	18.3732411	18.9120370	7.8976820
3	0.6140756	13.6820212	17.3090909	6.8832093
4	0.2093023	1.1337891	2.1864407	1.0480085
5	0.0000000	0.0000000	1.4285714	0.0000000
6	0.6620049	26.0889803	32.2090090	9.0479059
7	0.6822039	18.0978492	22.0266990	7.8463597
8	0.8495146	9.0171965	9.8095238	5.4617450
9	0.9768858	49.8262315	50.4136744	10.8386747
10	1.0057803	16.4220927	16.3894737	7.5139366
11	0.5764522	16.2571918	21.8571429	7.4792251
12	1.6313131	5.0307965	3.9600000	3.6904708
13	1.3416667	5.8527719	5.0526316	4.1143516
14	0.0000000	0.0000000	1.3333333	0.0000000
15	0.6470017	326.4070234	392.4015106	13.2879573
16	0.9540674	51.3783142	52.6004728	10.9103703
17	0.8609696	12.5069925	13.4885496	6.5725597
18	0.9062500	19.5984641	20.6192469	8.1157725
19	0.7786667	2.0962956	2.3885350	1.8207494
20	0.7493443	20.1269162	23.4310799	8.2049826
21	1.5994898	13.9597529	11.1266376	6.9527995
22	4.3478261	0.8426452	0.5750000	0.7943243
23	0.7034632	2.6781414	3.2307692	2.2442371
24	1.0836618	121.6824147	116.9764823	12.4361765
25	0.7542057	23.7855859	27.5226730	8.7539059
26	0.0000000	0.0000000	0.5000000	0.0000000
27	1.6538242	7.0092676	5.5796178	4.6541780
28	0.9530568	33.4658515	34.2874409	9.7968469
29	1.0130589	21.2835631	21.1470228	8.3908761
30	2.7882119	2.7443014	2.3828987	2.2905104
31	1.3089856	98.4840906	86.7784888	12.1438251
32	0.0000000	0.0000000	0.4948454	0.0000000
33	∞	0.0000000	0.0000000	0.0000000
34	0.6136364	0.6319149	0.8461538	0.6043450

각 논문의 OR과 대응되는 식 (6)의 w_i를 구하기 위해 먼저 n_{11i}, n_{12i}, n_{21i}, n_{22i}를 각각

대조군의 사망자 수, (대조군 환자 수 - 대조군 사망자 수), 실험군 사망자 수, (실험군 환자 수 - 실험군 사망자 수)으로 구한다. 예를 들면 논문번호 1번의 RCT 결과에 대한 n_{111}, n_{121}, n_{211}, n_{221}은 51, 151, 28, 176이 되고 $\widehat{OR}_1$은 $\frac{28\times151}{51\times176}$이고 w_1^{-1}는 $\frac{1}{51}+\frac{1}{151}+\frac{1}{28}+\frac{1}{176}$으로 추정이 된다. [표 4]는 각 논문별 $\widehat{OR}_i$, w_i, m_i, $w_i^{(R)}$을 계산한 결과표이다. 주목할 사항은 5, 14, 26, 32번 논문의 경우 OR이 0으로, 33번 논문은 ∞으로 추정이 되었다. 따라서 이들에 대한 log값은 정의가 되지 않거나 무한대이므로 식 (2)의 동질성 검정통계량인 Q의 계산에 사용되지 않고 따라서 Q통계량의 자유도는 33이 아닌 28이 된다. 식 (7)의 $\ln\widehat{OR}_{MH}$를 구하면 -0.1641이 되고 Q 통계량과 유의확률은 각각

$$Q=\sum_{i=1}^{34} w_i\left(\ln\widehat{OR}_i-\ln\widehat{OR}_{MH}\right)^2 \cdot I(i\notin\{5,14,26,32,33\})=86.07$$

과 1이 된다. 그러므로 동질성이 확보되지 않는다.

동질성이 확보되지 않으므로 랜덤효과 모형을 이용하여 메타분석을 실시한다. 이를 위해 식 (12)의 τ^2를 계산하면 다음과 같다.

$$\tau^2=(Q-\text{자유도})\times\left(\sum_i w_i-\frac{\sum_i w_i^2}{\sum_i w_i}\right)^{-1}=0.0722$$

랜덤효과 모형에 의한 OR과 신뢰구간을 추정하기 위해 식 (13)과 (14)는

$$\ln\widehat{OR}_{DL}=\frac{\sum_i w_i^{(R)}\ln\widehat{OR}_i}{\sum_i w_i^{(R)}}=-0.1081$$

$$\ln\widehat{OR}_{DL}\pm z_{\alpha/2}\left(\sum_i w_i^{(R)}\right)^{-1/2}=-0.1081\pm0.0713\cdot z_{\alpha/2}$$

이 되므로 각각 0.8976과 95% 신뢰구간은 (0.78, 1.03)이 된다.

이러한 결과를 forest plot을 이용하여 [그림 2]에 나타내었다. Forest plot은 각 연구들의 결과와 이를 바탕으로 한 메타분석의 결과를 종합하여 한 눈에 볼 수 있게 해주는 효과적인 그래프이다. [그림 2]를 보면 메타분석에 사용된 각 연구들의 $\exp(\hat{\theta}_i)=\widehat{OR}_i$과 이에 대응하는 95% 신뢰구간, 이들을 요약한 $\exp(\hat{\theta}_{pooled})$ 즉 summary 값과 이의 95% 신뢰구간을 나타내고 있다. 각 연구의 표본크기는 $\widehat{OR}_i$을 나타내는 점의 상대적인 크기로 비교할 수 있을 뿐만 아니라 신뢰구간의 폭에 따라 비교도 가능하다. [그림 2]의 5, 14, 26, 32, 33번

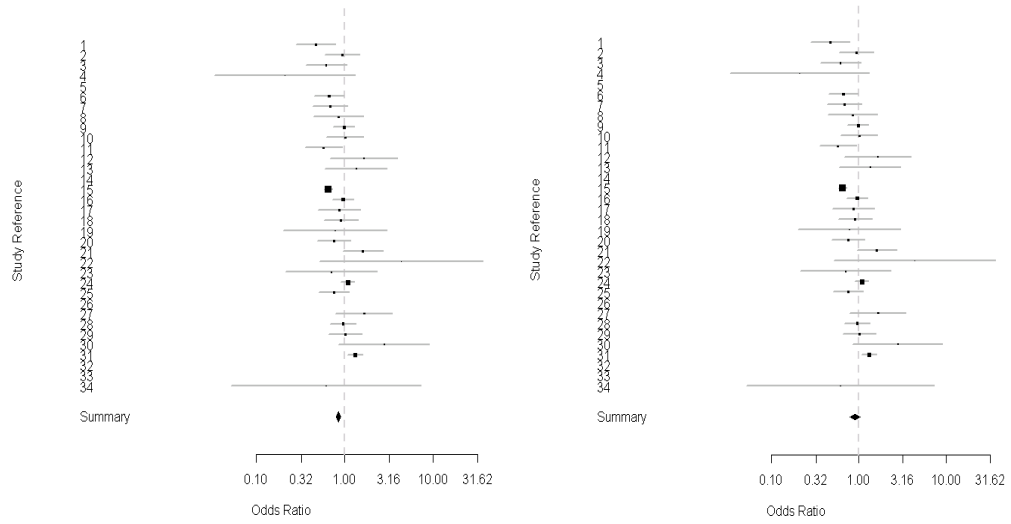

그림 2. Forest plot. 왼쪽과 오른쪽은 각각 모수효과 모형과 랜덤효과 모형을 사용하여 작성한 Forest plot이다.

논문의 경우에는 $\widehat{OR}_i$과 신뢰구간이 제시되지 않고 있다. 이 경우에는 앞에서도 언급했듯이, 각 논문의 OR이 0이거나 ∞로 추정되었기 때문이다.

모수효과 모형을 사용한 왼쪽의 forest plot에서는 summary 값으로 $\widehat{OR}_{MH}$을, 랜덤효과 모형을 사용한 오른쪽 그래프에서는 summary 값으로 $\widehat{OR}_{DL}$을 제시하고 있다. 각 연구 $\widehat{OR}_i$들의 동질성이 확보되지 않았으므로 랜덤효과 모형을 사용한 메타분석의 결과인 $\widehat{OR}_{DL}$을 보여주는 [그림 2]의 오른쪽 forest plot을 이용한다. 하지만 $\widehat{OR}_{DL}$의 폭으로 표현되는 신뢰구간이 1을 포함하고 있기 때문에 효과크기의 유의성은 확보되지 않음을 알 수 있다. 이는 매우 흥미로운 결과라 할 수 있다. 예를 들어 1번과 31번 논문 결과에 의하면 OR과

95% 신뢰구간이 각각 0.4710 (0.28, 0.78)과 1.3089 (1.07, 1.59)로 서로 상반된 결과였다. 이러한 상반된 결과는 이를 참고로 하는 연구의 진행에 혼란을 유발시킬 수 있다. 이에 반해 메타분석을 통해 OR과 95% 신뢰구간을 0.8976 (0.78, 1.03)로 구할 수 있고 따라서 유의한 효과가 없는 것으로 결론을 내릴 수 있기 때문에 이 주제에 관한 연구의 필요성은 적다라고 할 수 있겠다.

마지막으로, publication bias의 유무를 확인하기 위한 [그림 3]과 같은 funnel plot을 그

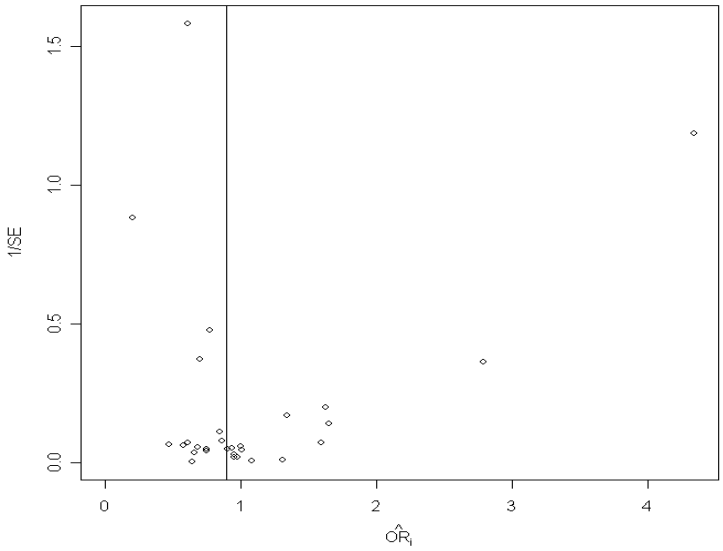

그림 3. Funnel plot. (vertical line: $\widehat{OR}_{DL}= 0.8976$)

려보았다. [그림 3]의 y축은 $\widehat{OR}_i$의 표본오차의 역수 즉, $1/w_i$로 하여 funnel plot을 작성하였고, 수직으로 그려진 선은 $\widehat{OR}_{DL}= 0.8976$일 때를 나타낸다. 이를 보면, $\widehat{OR}_{DL}$을 기준으로 대칭적인 분포를 이루지 않고 있다. 그러므로 메타분석의 결과에 publication bias가 내재될 가능성이 있다. 따라서 본 메타분석의 결론은 추가적인 논문 검색을 통해 보완한 후 결론을 내리는 것이 바람직하다.

참고문헌

[1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

[2] Begg, C.B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics,* 50, 1088-1101.

[3] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7, 177-188.

[4] Egger, M., Smith, G.D., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Br Med J,* 315, 629-634.

[5] Fisher, R.A. (1932). *Statistical Methods for Research Workers* (4th edn). Oliver and Boyd, London.

[6] Fleiss, J.L. (1993). The statistical basis of meta-analysis. *Stat. Methods Med. Res.,* 2, 121-145.

[7] Glass, G.V. (1976). Primary, secondary and meata-analysis of research. *Educational Researcher,* 5, 3-8.

[8] Greenland, S. and Robins J.M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics,* 41, 55-68.

[9] Hardy, R.J. and Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine,* 17, 841-856.

[10] Higgins, J.P.T. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine,* 21, 1539-1558.

[11] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute.* **22,** 719-748.

[12] Pearson, K. (1904). Report on certain enteric fever inoculations. *British Medical Journal,* **2**, 1243-1246.

[13] Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics,* **42**, 311-323.

[14] Rothman, K. J. and Boice, J. D. (1982). *Epidemiologic Analysis with a Programmable Calculator, 2nd edition*, Epidemiology Resources, Inc., Boston, MA.

[15] Smith, G.D., Song, F., Sheldon, T.A. and Song, F.J. (1993). Cholesterol lowering and mortality: The importance of considering initial level of risk. *Br. Med. J.*, **306**, 1367-1373.

[16] Tippett, L.H.C. (1931). *The Methods of Statistics*, Williams and Norgate, London.

[17] Whitehead, A. (2002). *Meta-analysis of Controlled Clinical Trials,* Wiley, Chichester.

[18] Woodword, M. (2005). *Epidemiology: Study Design and Data Analysis, 2nd edition*, Chapman & Hall,

Annals of Statistical Research
Vol. 4. No. 1, 2008, pp.51~57

On the Accuracy of Regression Analysis Using Microsoft Excel

Mi-Young Shin[1] Tae-Kyoung Cho[2]

Abstract

The numerical accuracy of regression analysis in Excel versions 2002 and 2007 is assessed. In particular, the previously indicated errors in Excel 2002, such as zero standard errors of the coefficients and the negative values of R^2, F and the sum of squares, have been eliminated in Excel 2007.

Keywords: numerical accuracy, regression, Microsoft Excel

1) Associate Professor, Department of Mathematics, The Catholic University of Korea, Bucheon-si, 420-743, Korea. E-mail: sati@catholic.ac.kr
2) Professor, Department of Statistics and Information Science, Dongguk University, Kyoungju, 780-814, Korea. E-mail: tkcho@dongguk.ac.kr

1. 서론

마이크로소프트 엑셀은 자료를 관리하고 분석하는데 사용하는 스프레드시트 프로그램이다. SAS나 SPSS와 같은 전문 통계패키지 만큼 통계분석 능력이 우수하지는 않으나 엑셀을 이용하여 회귀분석, 분산분석 등 기본적인 통계분석이 가능하며, 전문 통계패키지와 비교하여 저렴한 가격으로 인해 많은 기업과 개인들이 엑셀을 사용하고 있다. 또한 사용 방법의 간편성 때문에 많은 통계 관련 교재는 통계학의 개념을 설명하고 통계기법들을 실습하는 도구로서 엑셀의 사용법을 소개하고 있다.

그러나 엑셀의 통계분석 결과의 정확도에 대한 문제점은 그동안 많은 논문에서 지적되어 왔다. Cryer(2001)는 통계학의 여러 기본영역에서 엑셀의 심각한 문제점을 지적하며 통계분석에 엑셀을 사용해서는 안 된다고 주장하였다. Knusel(1998, 2004)은 엑셀 97에서 이항분포, 포아송분포, 정규분포 등 통계분포의 확률값에 오류가 있음을 찾아내었으며 엑셀 2003에서는 엑셀 97의 일부 오류는 수정되었으나 새로운 오류가 있음을 보고하였다. Yalta(2008)는 수리적 예를 통해 엑셀 2007도 다양한 통계분포의 계산에서 정확성과 안정성이 결여되어 통계분석에 사용하기에는 안전하지 않음을 보였다. McCullough와 Wilson (1999, 2008)은 추정, 난수생성, 통계분포의 확률값과 같은 세 영역에서 엑셀 97과 엑셀 2007의 신뢰성은 충분하지 않음으로 통계분석에 엑셀을 사용하지 말 것을 권하였다. Simonoff (2008)는 엑셀 2002와 Minitab을 이용한 회귀분석 결과를 비교하여 엑셀의 문제점을 설명하였다.

본 논문에서는 Simonoff (2008)가 사용했던 자료를 이용하여 Simonoff가 회귀분석에서 발견한 문제점들이 엑셀의 최근 버전인 엑셀 2007에서는 해결되었는지를 알아본다.

2. 회귀분석의 정확도에 대한 예제

Simonoff(2008)는 아래와 같은 자료를 사용해서 엑셀 2002의 회귀분석 결과와 Minitab의 회귀분석 결과를 비교해서 엑셀의 문제점을 설명하였다.

X (독립변수)	Y (종속변수)
10000000001	1000000000.000
10000000002	1000000000.000
10000000003	1000000000.900
10000000004	1000000001.100
10000000005	1000000001.010
10000000006	1000000000.990
10000000007	1000000001.100
10000000008	1000000000.999
10000000009	1000000000.000
10000000010	1000000000.001

위의 자료에 대한 Minitab과 엑셀 2002의 회귀분석 결과는 각각 다음과 같다.

<Minitab의 회귀분석 결과>

The regression equation is
Y=9.71E+08 + 0.0029 X

Predictor	Coef	StDev	T	P
Constant	970667056	616256122	1.58	0.154
X	0.00293	0.06163	0.05	0.963

S = 0.5597 R-Sq = 0.0% R-sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.0007	0.0007	0.00	0.963
Residual Error	8	2.5065	0.3133		
Total	9	2.5072			

<엑셀 2002의 회귀분석 결과>

요약 출력

회귀분석 통계량	
다중 상관계수	65535
결정계수	-0.538274369
조정된 결정계수	-0.730558665
표준 오차	0.694331016
관측수	10

분산분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	-1.34956254	-1.34956254	-2.79936728	#NUM!
잔차	8	3.85676448	0.48209556		
계	9	2.50720193			

	계수	표준 오차	t 통계량	P-값
Y 절편	2250000001	0	65535	#NUM!
X1	-0.125	0	65535	#NUM!

엑셀 2002의 결과를 살펴보면 회귀직선의 기울기가 음수로 잘못 계산되었으며, 음수가 될 수 없는 결정계수, 제곱합, 그리고 F비 값들이 음수로 나타났다. 또한 회귀계수의 표준오차를 0으로 계산하였으며, 따라서 t 통계량값도 정확한 값이 아니다.

Simonoff(2008)가 사용한 자료에 대해 엑셀의 최신 버전인 엑셀 2007의 회귀분석의 결과는 아래와 같다.

<엑셀 2007의 회귀분석 결과>

요약 출력

회귀분석 통계량	
다중 상관계수	0.016826509
결정계수	0.000283131
조정된 결정계수	-0.124681477
표준 오차	0.559742359
관측수	10

분산분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	0.000709868	0.000710986	0.002265693	0.963202
잔차	8	2.506492072	0.313311509		
계	9	2.507201939			

	계수	표준 오차	t 통계량	P-값
Y 절편	970666647.6	616256055.9	1.575102814	0.153881025
X1	0.002933335	0.061625606	0.047599294	0.963202431

위의 결과를 엑셀 2002의 결과와 비교해보면 회귀진선의 기울기로 양수로 수정되었으며, 음수가 될 수 없는 결정계수, F비, 제곱합과 같은 통계량의 값이 양수로 수정된 것을 알 수 있다. 엑셀 2002에서 발생한 회귀계수의 표준오차가 0이었던 오류와 t 통계랑의 오류도 모두 수정되었다. Minitab의 결과와 비교해보면 엑셀 2007에서 추정된 절편의 값만 Minitab의 값과 차이가 있고 나머지 통계량들은 모두 같은 것으로 나타나고 있다.

3. 결론

통계학을 이용한 자료분석에서 통계소프트웨어의 사용은 필수적이라 할 수 있다. 엑셀은 가장 많이 사용하는 소프트웨어 중 하나이나 엑셀의 통계분석 결과의 정확도에 대해 그동안 꾸준히 문제 제기가 있었으며, 엑셀을 통계분석도구로 사용하지 말 것을 권하는 연구결과도 있었다. 그럼에도 불구하고 많은 통계교재에서는 엑셀을 이용한 실습과 분석방법을 소개하고 있으며, 이러한 교재로 통계수업을 한 학생들은 엑셀이 갖고 있는 문제점을 인지하지 못한 채 엑셀을 통계분석 도구로 사용하고 있다. 엑셀은 사용방법이 간편하며 개인도 쉽게 구입할 수 있는 장점이 있어 통계 개념을 이해하고 자료분석 방법을 실습하는 도구로써 앞으로도 계속 사용하게 될 전망이다.

예제 자료를 이용하여 회귀분석의 정확도를 살펴본바 엑셀 2002에서 얻어진 회귀분석 결과의 많은 오류들이 엑셀 2007에서는 수정된 것을 알 수 있었다. 그러나 Yalta(2008)의 연구에서 알 수 있듯이 엑셀 2007에도 여전히 계산의 정확도에 문제가 있음을 알 수 있다. 이러한 문제점을 바로 알고 엑셀을 통계분석도구로 사용하는 것에 주의를 기울여 할 것이다.

참고문헌

[1] Cryer, J. D. (2001). Problems with using Microsoft Excel for Statistics. *Proceedings of 2001 the Joint Statistical Meetings*.

[2] Knusel, L. (1998). On the Accuracy of Statistical Distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis*, Vol 26, 375-377.

[3] Knusel, L. (2004). On the accuracy of statistical distributions in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, Vol 48, 445-449.

[4] McCullough, B. D. and Wilson, B. (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis*, Vol 31, 27-37.

[5] McCullough, B. D. and Heiser, D. A. (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics and Data Analysis*, Vol 52, 4570-4578.

[6] Simonoff, J. (2008). Statistical analysis using Microsoft Excel. Available from: http://www.stern.nyu.edu/~jsimonof/classes/1305/pdf/excelreg.pdf

[7] Yalta, A. T. (2008). The accuracy of statistical distributions in Microsoft Excel 2007. *Computational Statistics and Data Analysis*, Vol 52, 4579-4586.