

## 학년진급률에 따른 학생수 예측방법

김중태<sup>1</sup>

<sup>1</sup>대구대학교 전산통계학과

접수 2009년 7월 13일, 수정 2009년 9월 16일, 게재확정 2009년 9월 21일

### 요약

본 연구는 학년(연령) 진급에 따른 인구증감률에 대하여 전국 학생수를 예측하는 다양한 방법들을 제시하고, 제시된 예측 모형들을 이용하여 출생아들이 고3학생이 되는 18년 후인 2026까지의 학생수를 예측하는 것이다. 이동평균과 시계열모형, 회귀분석 등 다양한 예측모형들이 사용되었고, 적합척도들을 이용하여 이들의 오차들을 측정하였다. 예측오차를 측정하는 도구들을 기준으로 제시된 예측방법들 중 이동평균에 의한 방법은 쉽고 단순한 장점을 지니면서도 기존에 예측되어진 한국교육개발원의 예측결과 뿐 아니라 회귀분석 및 시계열예측의 고등기법들의 결과들 보다 예측 능력이 우수한 것으로 나타났다.

주요용어: 시계열분석, 이동평균, 인구통계, 홀트-윈터스모형, 회귀분석.

### 1. 서론

교육과학기술부 소속인 한국교육개발원(2007)은 교육통계연구센터의 교육통계서비스 시스템을 통해서 초·중·고등학교 학생수를 2019년까지 예측하였다. 한국교육개발원의 예측 결과를 보면 2016년에서 2017년의 1년 사이에 고3학생수는 무려 10만 명이상이 감소한다고 예측하였다. 이러한 한국교육개발원의 학생수 예측결과는 전국의 초·중·고등학교들뿐만 아니라 학생수급이 원활하지 않는 중소 대학들의 학생수급에 대한 정책수립에 심각한 영향을 미친다.

그러므로 학생(연령)수에 대한 예측은 학문적인 가치보다는 행정 정책의 결정에 매우 중요한 영향을 미친다. 측정 가능한 장래 인구수를 예측하기 위해서는 최근 출생아수의 데이터가 필수적인 요소이다. 즉, 최근 출생아수는 18년 후인 2026년에 장래 고3학생수를 예측할 수 있는 중요한 변수이고, 출생아수의 분포와 고3학생수의 분포와는 밀접한 관계를 가진다. 그러므로 2026년의 고3학생수는 2008년의 0세인구로 추정이 가능하지만, 2027년 이후의 고3학생수를 정확히 예측하기 위해서는 2009년 이후의 출생아수를 예측하여야만 가능하다. 그러나 미래에 태어날 출생아수의 예측은 쉽지 않고, 출생아수는 사회적인 환경과 정부의 출산장려 정책 등, 여러 가지 변수에 따라서 영향을 받을 것이다.

본 연구에서 학년(연령) 진급에 따른 인구 증감률 추정모형을 기반으로 최근의 출생아수와 각 연령별 데이터들을 이용하여 출생아들이 고3학생이 되는 18년 후인 2026까지의 전국 초·중·고등학교 학생수를 예측하는 다양한 예측모형을 제시하였다. 사용된 예측모형으로는 이동평균, 회귀모형, 시계열 모형의 선형이동과 홀트-윈터스(Holt-Winters) 모형, SARIMA 모형 등이다.

기존의 초·중·고등학교의 학생수에 대한 연구로 한국교육개발원(2005, 2006, 2007)이 제공한 교육예측통계시스템에서 초·중·고등학교의 학생수를 2019년까지 예측하였는데, 예측 모형으로 시계열 선형모

<sup>1</sup> (712-714) 경상북도 경산시 진량면 내리동15, 대구대학교 전산통계학과, 교수. E-mail: jtkim@daegu.ac.kr

형을 가정하고, 선형모형의 모수를 추정하기 위하여 로지스틱성장곡선함수와 로지스틱 지수평활함수를 사용하였다. 그러나 시계열 예측에 있어서 로지스틱 함수를 적용할 경우에 초기값에 큰 영향을 받는 것으로 알려졌고 (김연형, 1994), 실제로 비선형모형함수를 사용하여 예측한 결과 예측력이 매우 떨어짐을 발견하였다.

Kim (2005 a, b)는 이동평균을 이용한 예측방법으로 2005년에서 2022년까지 대구·경북의 초·중·고등학교의 학생수에 대한 예측과 고3학생수 대비 대학 입학정원과의 관계를 다루었다. 김종태 등 (2009)에서는 2009년에서 2026년까지 대구·경북의 초·중·고등학교의 학생수에 대한 예측과 고3학생수 대비 대학 입학정원과의 관계를 분석하였다. 김종태 (2009)에서는 학년의 진급에 따라서, 초등1학년에서 고3학년 까지 12년 한 주기로 하는 시계열 데이터 전환과, 각 학년을 한 주기로 하는 시계열 데이터 전환방법을 사용하여, 시계열 모형인 Holt-윈터스 가법모형을 이용하여 학생수를 예측하였다. 조찬혁 (2008)과 BakiBillah 등 (2006)의 연구에 따르면, Holt-윈터스 (Holt-Winters) 모형은 기법의 간명성과 단순성에도 불구하고 예측능력 만큼은 Box-Jenkins의 ARIMA와 같은 모형들에 결코 뒤지지 않는다고 평가하고 있다.

2절에서는 학년 (연령)진급에 따른 인구의 표기 방법을 정의하고, 그에 따른 인구 증감률을 정의하였다. 3절에서는 비례법,  $m$ 이동평균법,  $m \times n$ 이동평균법, Holt-윈터스 가법모형, SARIMA모형, 회귀분석모형을 이용한 예측방법을 설명하였다. 4절에는 모의실험을 사용하여 제시된 모형들에 대한 적합척도들을 구하고, 각 모형에 따른 2026년까지의 학생수를 예측하고, 한국교육개발원 (2007)의 교육예측통계시스템에 있는 고3학생수 예측결과와 비교 분석하였다

## 2. 학년 (연령)진급에 따른 인구 증감률 정의

$X_{d,y}$ 를  $y$ 연도의  $d$  ( $d = 0, 2, \dots, 18$ )인구수로 다음과 같이 정의하자.

$$d = \begin{cases} 0, 1, \dots, 6 & \text{각각 0세, 1세, } \dots, 6\text{세 인구수,} \\ 7, 8, \dots, 12 & \text{각각 초등 1, 2, } \dots, 6\text{학년 학생수,} \\ 13, 14, 15 & \text{각각 중학 1, 2, 3학년 학생수,} \\ 16, 17, 18, & \text{각각 고등 1, 2, 3학년 학생수.} \end{cases} \quad (2.1)$$

여기서 0세 - 6세 인구는 주민등록인구를 사용하였다. 비록 주민등록 인구데이터는 0세에서 7세가 될 때까지 인구수가 꾸준히 증가하는 모순을 가지고 있다. 그러나 이러한 모순에도 불구하고, 통계청 (2006)의 장래인구추계 보다는 주민등록인구를 사용한 이유는 각 연령층의 증가와 감소가 매우 안정적이기 때문이다.

초·중·고등학교의 학생수에 대한 통계는 한국교육개발원 (1982 - 2008)의 '교육통계연감'에서 발표한 학생수 자료를 이용하였다.

위의 식 (2.1)에 대한 예를 들면, 2008년도의 0세 인구,  $X_{d,y} = X_{0,2008}$ 는 18년 후인 2026년에는 고3학생수,  $X_{d+18,y+d+18} = X_{18,2026}$ 가 된다. 학년 (연령)진급에 따른 연도의 이동을 표로 만들어 보면 다음과 같다.

1991년 0세가 18년 후인 2009년 고3학생이 되고, 1992년 0세는 17년 후인 2009년에 고2학생, 18년 후인 2010년에는 고3학생이 된다. 이런 규칙에 따라서 2008년 0세는 1년 후인 2009년에는 1세, 2010년에는 2세,  $\dots$ , 18년 후인 2026년에는 고3학생이 된다. 학년 (연령)인구 증감률은 다음과 같이 정의 된다.

$$P_{(d+1,y+d+1)} = \frac{X_{d+1,y+d+1}}{X_{d,y+d}}. \quad y = 1982, 1983, \dots, 2008. \quad (2.2)$$

표 2.1 출생연도를 기준한, 학년 (연령)진급에 따른 연도 이동표

연령	출생연도			
	1982년	1983년	...	2008년
0세	$X_{0,1982}$	$X_{0,1983}$	...	$X_{0,2008}$
1세	$X_{1,1983}$	$X_{1,1984}$	...	$X_{1,2009}$
2세	$X_{2,1984}$	$X_{2,1985}$	...	$X_{2,2010}$
...	...	...	...	...
고3	$X_{18,2000}$	$X_{18,2001}$	...	$X_{18,2026}$

식 (2.2)에서  $d(d = 0, 1, \dots, 17)$ 는 식 (2.1)에 정의 된 것으로  $d + 1 \leq 18$ 이다.

표 2.1에서 나타낸 것 같이, 0세 인구가 시작되어 고3 인구가 되는 19년 동안을 한 주기로 볼 때,  $y(y = 1982, 1983, \dots, 2008)$ 는 0세 인구가 시작되는 연도를 의미한다.  $X_{d,y+d}$ 는  $y + d$ 년도에  $d$ 연령의 인구수이고,  $X_{d+1,y+d+1}$ 는  $y + d + 1$ 의 년도에  $d + 1$ 연령의 인구수이다. 예를 들어,  $y = 1982$ 일 때,  $d = 0$ 이면,  $X_{d,y} = X_{0,1982}$ 는 1982년의 0세 인구이다. 이들 0세 인구가 18년 후에는  $d = 18$ 이 되고,  $y + d = 2000$ 에 고3이 되고,  $X_{d,y+d} = X_{18,2000}$ 은 1982년의 0세인구가 2000년에 고3이 되는 인구수이다. 실제로 1982년의 고3학생수는 1982년으로부터 18년 전인 1964년 ( $y = 1964$ )에 0세 인구수가 출발한 것이고, 1982년의 고2학생수는 1965년 ( $y = 1965$ )에 0세 인구수에서, 출발된 것으로 1983년에는 고3학생수가 된 것이다. 그러면 식 (2.2)에서 제시한  $P_{(d+1,y+d+1)}$ 는  $y + d$ 년도에서  $y + d + 1$ 년 사이에 한 해 동안의 인구증감률이 된다.

식 (2.2)에 의한 인구 증감률에 대한 분포를 최근 10년간 (1999년 - 2008년)을 그래프로 나타내면 다음과 같다.

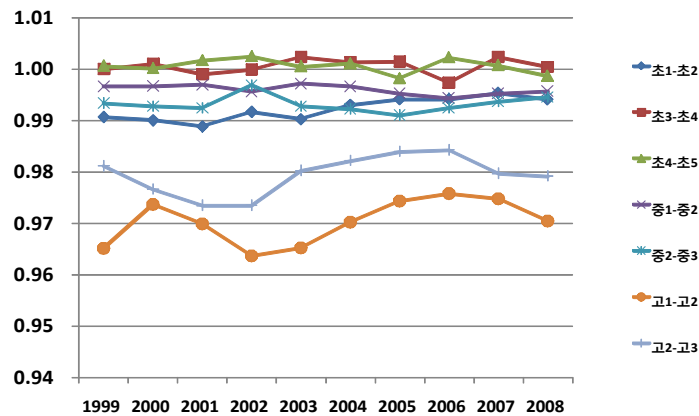


그림 2.1 초·중·고등학교의 진급률에 대한 분포

### 3. 제시된 장래 학생수 예측 방법

$X_{d+1,y+d+2}$ 를 다음 해의 알지 못하는 (미지의)  $d+1$ 학년 (연령)의 인구수라고 가정하자.  $d+1$ 학년 (연령)의 미지의 인구 증감률  $P_{(d+1,y+d+2)}$ 을 추정함으로써  $X_{d+1,y+d+2}$ 는 다음과 같이 추정할 수 있다.

$$\hat{X}_{d+1,y+d+2} = \hat{P}_{(d+1,y+d+2)} \cdot X_{d,y+d+1}. \quad (3.1)$$

식 (2.2)의 학년 (연령) 인구 증감률  $P_{(d+1,y+d+2)}$ 의 추정치  $\hat{P}_{(d+1,y+d+2)}$ 을 구하기 위한 방법으로 다음의 모형들을 고려한다.

#### 3.1. 모형 A: 인구 증감률 비례법에 의한 추정

그림 2.2에서 진급률의 분포를 살펴보면 어느 정도 일정한 패턴을 가지고 있는 것을 알 수 있다. 이러한 관점에 직전 연도의 인구 증감률과 당해 연도의 인구 증감률이 같을 것이라는 가정 하에서 비례법을 이용하여 다음과 같이 인구를 추정할 수 있다. 즉,  $X_{d+1,y+d+1} : X_{d,y+d} = X_{d+1,y+d+2} : X_{d,y+d+1}$ 라 가정하면, 증감률의 추정치  $\hat{P}_{(d+1,y+d+2)}^A$ 은 다음과 같다.

$$\hat{P}_{(d+1,y+d+2)}^A = P_{(d+1,y+d+1)}. \quad (3.2)$$

식 (3.2)의 추정된 증감률을 이용하여, 미지의 학년 (연령) 인구수  $X_{d+1,y+d+2}$ 는 다음과 같이 추정된다.

$$\text{모형 A: } \hat{X}_{d+1,y+d+2} = P_{(d+1,y+d+1)}^A \cdot X_{d,y+d+1}. \quad (3.3)$$

그러나 직전 연도의 증감률을 당해 연도의 증감률로 사용한다는 것은 매우 강한 가정이다. 그럼에도 불구하고 모형 A를 사용하는 이유는 예측력은 떨어지지만 추정이 방법이 매우 쉽고, 단순하기 때문이다.

#### 3.2. 모형 B: 인구 증감률에 대한 $m$ 이동평균에 의한 추정

3.1절의 모형 A를 보완하기 위해서, 직전 연도의 증감률을 포함하는 최근  $m$ 개의 증감률에 대한 이동평균을 다음 연도의 증감률의 추정치로 사용하는 것이다.  $m$ 이동평균 방법은 쉽고, 단순하면서도 우수한 예측력을 지니고 있다.  $m$ 이동평균에 대한 미지의 학년 (연령) 인구수  $X_{d+1,y+d+2}$ 에 대한 모형 추정식은 다음과 같다.

$$\hat{P}_{(m, d+1,y+d+2)}^B = \frac{1}{m} \sum_{k=1}^m P_{(d+1,y-k+d+2)}. \quad (3.4)$$

$$\text{모형 B: } \hat{X}_{m, d+1,y+d+2} = \hat{P}_{(m, d+1,y+d+2)}^B \cdot X_{d,y+d+1}. \quad (3.5)$$

여기서  $m$ 은  $d+1$ 연령의 최근 증감률을 포함해서 사용되는 증감률의 개수를 의미한다. 본 연구에서는  $m = 3, 4, 5, 6$ 을 사용하였다.

### 3.3. 모형 C: 인구 증감률에 대한 $m \times n$ 이동평균에 의한 추정

3.2절의 모형 B를 보완하기 위해서, 식 (3.4)의 증감률  $m$ 이동평균  $\hat{P}_{(m,d+1,y+d+2)}^B$  들에 대하여  $n$ 개의 평균을 구하여 다음 연도의 증감률의 추정치로 사용하는 것이다.

$$\hat{P}_{(m \times n, d+1, y+d+2)}^C = \frac{1}{n} \sum_{k=1}^n \hat{P}_{(m, d+1, y-k+d+2)}^B. \quad (3.6)$$

$$\text{모형 C: } \hat{X}_{m \times n, d+1, y+d+2} = \hat{P}_{(m \times n, d+1, y+d+2)}^C \cdot X_{d, y+d+1}. \quad (3.7)$$

모형 B와 모형 C에서 이동평균의 정의는 시계열에서의 이동평균의 정의와는 다소 차이가 있다. 시계열에서의 이동평균은 데이터들의 중간값들을 추정하는 특성을 가지지만 모형 B와 모형 C의 이동평균의 개념은 가장 최근의 증감률에 대하여 가중치를 더해가는 특성을 가지고 있다.

이들  $m \times n$ 이동평균 방법은  $m$ 이동평균들 보다 더 예측오차가 적은 것으로 나타난다. 특히 모의실험 결과  $4 \times 4$ 이동평균의 학생수에 대한 예측 결과가 좋은 것으로 나타난다.

### 3.4. 모형 D: 홀트-윈터스 가법모형을 이용한 인구수 추정

계절성을 가지는 시계열 추정을 위해서는 학생수 데이터를 시계열 데이터로 전환 시킨 후에, 2008년도 초등1학년의 학생수가 2019년에 고3학생수가 되기까지의 학생수에 대한 추정을 홀트-윈터스의 방법을 가지고 예측하였다. 홀트-윈터스 모형은 다음과 같다.

$$L_t = \alpha(Y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1}),$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1},$$

$$S_t = \gamma(Y_t - L_t) + (1 - \gamma)S_{t-s},$$

$$F_{t+m} = L_t + b_t m + S_{t-s+m}.$$

여기서  $Y_t$  = 시점  $t$ 에서 관찰된 값,  $L_t$  = 시점  $t$ 에서 시계열 평균수준,  $b_t$  = 시점  $t$ 에서 시계열 추세성분,  $S_t$  = 시점  $t$ 에서 시계열 계절성분,  $F_{t+m}$  = 시점  $t$ 에서 예측한 시점  $t+m$ 의 예측값,  $s$  = 계절성분의 길이, 그리고  $\alpha, \beta, \gamma$  = 평활모수이다.

모형 D는 김종태 (2009)의 추계에 대하여 2026년까지 확장한 것이다. 홀트-윈터스 모형을 사용하기 위해서는 1998년 이전의 많은 데이터가 필요하다. 이러한 이유로 0세 - 6세 인구를 1982년 - 2005년까지는 통계청 (2006) '장래인구추계' 데이터를 사용하였고, 장래추계인구 데이터에서 2006년 - 2008년까지는 예측 데이터이므로 실제 데이터인 주민등록인구 데이터를 사용하였다.

$$\text{모형 D1: } \hat{X}_{d+1, y+d+2} \equiv F_{t+1}. \quad (1\text{주기 당 } 19\text{개 데이터}). \quad (3.8)$$

모형 D1은 표 2.1에서 세로줄인 0세에서 고3 (19개)까지를 하나의 주기로 하여 시계열 데이터를 생성하여, 미지의 학생수를 예측하였다.

$$\text{모형 D2: } \hat{X}_{d+1, y+d+2} \equiv F_{t+1}. \quad (1\text{주기 당 } 27\text{개 데이터}). \quad (3.9)$$

모형 D2는 표 2.1에서 가로줄의 각각의 연령층에 대하여 27개의 데이터를 하나의 주기로 하여 시계열 데이터를 생성하여, 미지의 학생수를 예측하였다.

모형 D1과 D2에서 평활모수  $\alpha, \beta, \gamma$ 의 값은 절대오차의 평균을 가장 작게 하는 값들을 구하여 사용하였다.

### 3.5. 모형 E: Holt - SARIMA (p, d, q)(P, D, Q)모형을 이용한 인구수 추정

시계열 분석 모형인 SARIMA (p, d, q)(P, D, Q)모형을 이용하여 예측을 하였다. 그러나 증감률을 추정하기 위하여 시계열 모형의 ARIMA모형이나 평활모형 등을 사용할 경우에 각 연령별 계층 (19개 계층)들에 대한 증감률의 분포 변화에 따라서, 그리고 각 분포들의 연도별 변화에 따른 분포의 이동 범위 등에 따라 각 모수들을 추정해야 하는 번거로움과 복잡성을 가지고 있다.

### 3.6. 모형 F: 회귀분석을 이용한 인구수 추정

출생이수와 장래인구추계 0세 인구를 독립변수로 하여 각 학년 (연령)별 인구수에 대한 단순회귀분석으로도 예측이 가능하다.

$$\text{모형 F1: 각 학년 (연령)의 인구수}_i = \alpha + \beta(\text{출생이수}_i) + \epsilon_i, \quad (3.10)$$

$$\text{모형 F2: 각 학년 (연령)의 인구수}_i = \alpha + \beta(\text{장래인구추계 0세 인구}_i) + \epsilon_i. \quad (3.11)$$

위의 모형 F1과 F2에서 각 학년 (연령)별 인구수는 출생이수 혹은 장래인구추계 0세 인구의 선형 상관관계를 가지는 있는 것에 기초한 것이다.

### 3.7. 선형추세와 비선형 추세 방법들

선형추세를 보이는 시계열을 예측하는 방법으로 선형이동평균법, 선형추세법, 이중지수평활법, Holt의 선형지수 평활법 등이 있다. 그러나 이들 방법으로 예측한 결과들은 모의실험 결과 위에서 제시한 모형들 보다 예측력이 매우 낮았다.

비선형추세의 평활법들인 지수곡선, 지수성장 곡선, 로지스틱 곡선 로그 곡선 등의 모형들 역시 모의 실험 결과 위에서 제시한 모형들 보다 예측력이 매우 낮았다.

## 4. 학생수 추정오차와 예측결과와 결론

3절의 모형A에서 모형 E까지의 예측값과 실제값들의 오차를 평가하는 기준으로 다음과 같은 적합척도들인, 평균절대편차 (Mean Absolute Deviation; MAD), 제곱근평균제곱오차 (Root Mean Squared Error; RMSE), 평균절대백분비오차 (Mean Absolute Percentage Error; MAPE), 오차평균 (Mean of Error; ME)을 사용한다.

$$\text{MAD} = \frac{\sum_{t=1}^n |Y_t - \hat{Y}_t|}{n}, \quad \text{RMSE} = \sqrt{\frac{\sum_{t=1}^n (Y_t - \hat{Y}_t)^2}{n-1}},$$

$$\text{MAPE} = \frac{\sum_{t=1}^n \left( \frac{|Y_t - \hat{Y}_t|}{Y_t} \right) \times 100}{n}, \quad \text{ME} = \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t).$$

표 4.1에서는 적합척도인, 평균절대편차 (MAD), 제곱근평균제곱오차 (RMSE), 평균절대백분비오차 (MAPE), 오차평균 (ME)를 기준으로 모의실험을 통해서 제시된 모형들의 추정값들과 실제값의 오차들을 비교하여 예측 능력을 조사하였다.

표 4.1에서 모형 A의 비례법은 다른 모형들보다 상당히 큰 오차들을 가진다. 오차평균 (ME)를 기준으로 할 때, 모형 B의 5이동평균이 가장 적은 값을 가지고, 5×5이동평균, 5×4이동평균, 4이동평

균, 4×4이동평균 순으로 점점 커진다. 그러나 평균절대편차 (MAD), 제곱근평균제곱오차 (RMSE), 평균절대백분비오차 (MAPE)들을 기준으로 볼 때에는 모형 C의 4×4이동평균이 오차값들이 가장 적고, 4×3이동평균, 4이동평균, 5×4이동평균, 5×5이동평균 순으로 적합척도 값들이 점점 커짐을 알 수 있다. 시계열 분석 모형인 Holt-윈터스 모형 D와 SARIMA 모형 E는 모든 적합척도의 결과들에서 이동평균 모형들과 비교할 때, 상당히 큰 오차를 가지는 것으로 나타난다.

표 4.1 모형 A에서 모형 E에 대한 적합척도 비교

모형		ME	MAD	MAPE	RMSE	SSE	MSE
모형 A	비례법	13,582	15,022	2.207	21,731	5.15E+10	4.72E+08
모형 B	3이동평균	1,457	4,937	0.773	6,879	5.16E+09	4.73E+07
	4이동평균	792	4,799	0.749	6,335	2.62E+09	2.40E+07
	5이동평균	712	4,954	0.774	6,550	2.90E+09	2.66E+07
	6이동평균	1,097	5,218	0.818	6,974	3.35E+09	3.07E+07
모형 C	3×3이동평균	1,176	5,151	0.808	6,796	3.12E+09	2.86E+07
	4×3이동평균	894	4,786	0.749	6,256	2.52E+09	2.31E+07
	4×4이동평균	829	4,757	0.744	6,213	2.48E+09	2.28E+07
	5×4이동평균	778	4,901	0.767	6,387	2.70E+09	2.48E+07
	5×5이동평균	749	4,928	0.770	6,488	2.81E+09	2.58E+07
	6×5이동평균	1,054	5,103	0.800	6,779	5.01E+09	4.60E+07
모형 D	19주기	5,641	7,597	1.102	8,614	8.09E+09	7.42E+07
	27주기	5,279	7,276	1.079	8,566	8.00E+09	7.34E+07
모형 E	SARIMA	-4,917	6,012	0.912	7,828	6.68E+09	6.13E+07

모형 D와 모형 E의 시계열 분석 방법인 Holt-윈터스 가법모형이나 SARIMA (p, d, q)(P, D, Q)모형은 각각의 적합한 모수들을 찾는데 시간이 많이 걸릴 뿐만 아니라 모형의 가정들을 충족시키기 위한 문제점들을 수반한다. 장래의 인구를 예측하는데 있어서 시계열 분석 방법들인 모형 D와 모형 E는 이동평균에 의한 방법들인 모형 B와 모형 C 보다 적합척도들의 값들이 좋지 않은 결과를 나타내고 있다.

모형 F의 회귀분석을 이용한 인구수추정은 다른 모형들과 비교해 볼 때, 측정 오차들이 매우 크게 추정되는 결과를 가져오기에 다음의 표 4.1에서 생략하였다. 모형 F의 회귀분석에 의한 미래 학년 (연령) 인구수의 예측 역시 각 연령 또는 학년에 따른 회귀모형들이 달라질 뿐만 아니라, 모든 학년 변수들을 사용할 경우에 매번 적절한 변수를 선택해야 하는 번거로움을 지닌다. 식 (3.10)의 모형 F1은 예측력이 과대추정이 되고, 모형 F2는 과소추정이 된다.

표 4.2는 한국교육개발원 (2007) 학생수 예측에 대한 것이고, 표 4.3은 모형 C의 4×4이동평균에 의한 예측 값이다. 다른 모형들의 예측값들은 생략하였다.

그림 4.1에서 모형 A는 비례법 결과이고, 모형 B에서는 4이동평균법의 결과이고, 모형 C는 4×4이동평균법의 결과이고, 모형 F는 식 (3.10)의 출생아수와 고3학생수에 대한 단순 회귀분석의 결과이고, 기존의 출생아수를 18년 평행이동 시킨 결과와 한국교육개발원 (2007) 학생수 예측을 비교한 것이다.

그림 4.1은 미래의 인구 예측에 가장 큰 영향을 미치는 데이터인 출생아수들을 예측된 시점으로 평행 이동 시킨 후에, 고3학생수에 대한 제시된 모형들의 예측값의 분포 모양의 비교를 함으로서 모형들의 타당성을 검토하기 위한 것이다.

모형 A, B, C, F는 출생아수의 분포와 거의 같은 분포의 모양을 가지고 있지만, 한국교육개발원 (2007) 학생수 예측 모형은 출생아수의 모형과 매우 다른 모습을 나타내고 있다. 모형 F의 출생아 수에 대한 고3학생수의 단순 선형회귀 추정 결과는 다른 모형들에 비해 과대 추정되는 결과를 보이고 있다. 현재의 출생아수와 분포는 18년 후의 고3학생수의 분포에 매우 많은 영향을 끼침으로서 두 개의 분포는 같은 모양을 나타내는 것이 타당하다고 본다. 그 이유는 어느 한 해에 태어난 출생아들이 자연발생

표 4.2 한국교육개발원 (2007) 학생수에 대한 예측 결과

	초등1	초등2	초등3	초등4	초등5	초등6	중등1	중등2	중등3	고등1	고등2	고등3
2009	474,946	486,910	606,227	597,716	616,808	644,835	645,284	673,774	672,856	661,550	671,095	647,431
2010	450,528	476,941	486,563	605,458	597,178	615,500	615,867	642,444	669,778	658,393	648,464	664,750
2011	429,560	452,784	476,593	485,889	604,858	595,806	596,041	613,171	638,681	627,696	646,099	642,992
2012	431,929	431,850	452,450	475,873	485,363	603,354	603,536	593,367	609,619	599,012	616,638	641,216
2013	432,811	434,367	431,525	451,714	475,320	484,067	484,161	600,678	589,973	579,587	589,083	612,427
2014	430,038	435,390	434,037	430,776	451,149	473,967	473,964	481,658	597,286	586,651	570,568	585,383
2015	426,066	432,738	435,050	433,230	430,199	449,775	449,732	471,187	478,971	470,396	578,103	567,208
2016	421,020	428,879	432,396	434,190	432,607	428,813	428,729	446,628	468,593	460,110	468,750	574,822
2017	415,089	423,933	428,538	431,495	433,530	431,139	430,968	425,121	444,199	436,068	454,234	466,013
2018	408,537	418,094	423,587	427,592	430,799	431,973	431,757	426,431	422,834	415,013	430,903	451,563
2019	401,724	411,495	417,756	422,653	426,901	429,250	428,994	427,214	424,137	416,205	410,099	427,950
2020	394,959	404,632	411,163	416,833	421,975	425,369	425,028	424,479	424,919	416,887	411,283	406,817
2021	388,327	397,819	404,304	410,253	416,165	420,460	420,081	420,559	422,196	414,131	411,953	407,857
2022	382,292	391,138	397,494	403,410	409,592	414,666	414,254	415,662	418,298	410,221	409,229	408,523

표 4.3 모형 C의 4 × 4 이동평균 학생수에 대한 예측 결과

	초등1	초등2	초등3	초등4	초등5	초등6	중등1	중등2	중등3	고등1	고등2	고등3
2009												634475
2010											663426	650450
2011										668480	648539	635982
2012									673083	659548	639831	627679
2013								676498	672083	658928	639716	627448
2014							653963	651130	646971	634345	615664	603799
2015						640371	641843	639177	635071	622576	604159	592558
2016					611728	608977	610604	608024	604097	592185	574703	563691
2017				593231	590076	587673	588984	586494	582711	571279	554467	543820
2018			599719	596844	593894	591347	592717	590207	586412	574900	557943	547225
2019		529634	527145	524757	522132	519883	521120	518926	515584	505447	490533	481118
2020	472631	470894	468820	466702	464335	462343	463452	461494	458520	449506	436252	427879
2021	469486	467869	465806	463673	461316	459360	460431	458486	455533	446585	433420	425098
2022	451501	449956	447944	445881	443640	441741	442782	440912	438073	429465	416800	408797
2023	415433	413990	412129	410250	408180	406433	407395	405675	403062	395141	383488	376126
2024	427559	426061	424167	422231	420096	418300	419289	417518	414829	406677	394686	387109
2025	470945	469316	467224	465086	462736	460760	461845	459895	456933	447955	434746	426400
2026	446091	444546	442560	440534	438312	436437	437468	435620	432815	424310	411797	403892

적인 감소를 제외하고는 갑자기 한 연령대의 인구수가 격감하지 않기 때문이다.

출생아수의 연령이동 분포와 한국교육개발원 (2007)의 예측결과 분포를 비교해 보면, 2017년부터 2021년까지의 고3학생수에 대한 분포는 다르게 보인다. 즉, 18년 후로 연령이동 출생아수는 2017년까지 서서히 감소하다가 2018년에 증가하고 난 다음에 2019년과 2020년까지 가파르게 감소한 후에 점진적인 변화가 일어난다.

그러나 한국교육개발원 (2007)의 고3학생수 결과는 2016년에서 2017년 사이에 급격히 변화하고, 2022년까지 제시된 다른 모형들 보다 과소 추정된 결과들을 보인다. 이는 0세에서 6세의 인구자료를 본 연구에서는 주민등록자료를 사용하였고, 한국교육개발원 (2007)의 고3학생수 추정에서는 통계청의 장래인구추계 자료를 사용한 것으로 추정된다. 장래인구추계자료는 추계에 있어서 정교하게 했음에도 불구하고, 결과적으로는 상당히 많은 문제점을 가지고 있는 것으로 조사되고 있다. 이 부분에 대한 언급은 향후의 연구에서 다룰 것이다.



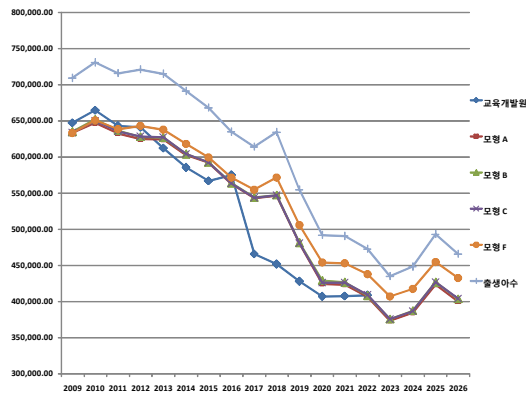


그림 4.1 출생아 수를 기준으로 한 고3학생수에 대한 예측 결과 비교

결론적으로 표 4.1의 적합적도와 그림 4.1의 결과들을 비교해 볼 때, 모형 C의 4×4 이동평균법의 예측결과가 가장 우수한 것으로 나타난다.

다음 표 4.4는 고3학생수에 대한 교육개발원 예측결과와 모형 C의 결과를 비교한 것이다. 두 모형 간의 예측의 차이는 2017년에서 2019년 사이에서 일어난다. 두 모형 사이의 차이는 2017년에 77,807명, 2018년에 95,662명, 2019년에 53,168명의 차이를 나타내고 있다.

2012년부터 현재 2009년의 고3학생수 보다 감소하기 시작하여, 교육개발원은 2016년 - 2017년 사이에 108,809명의 고3학생수가 줄어든다고 예측하는 반면에 모형 C에서는 2018년 - 2019년 사이에 66,107명, 2019년 - 2020년 사이에 53,239명의 고3학생수가 감소할 것으로 예측한다. 그러나 2022년에 이르러서는 교육개발원의 고3학생수 예측이나 모형 C의 고3학생수 예측은 큰 차이를 나타내지 않는다.

결론적으로 교육개발원의 예측에 따르면 2017년에 고3학생수가 십만 명이상이 감소하고, 모형 C에 따르면 2019년과 2020년에 6만 명과 5만 명이상이 감소한다고 예측된다.

표 4.4 고3학생수에 대한 교육개발원 예측결과와 모형 C 결과 비교

연도	2009	2010	2011	2012	2013	2014	2015	2016	2017
교육개발원	647,431	664,750	642,992	641,216	612,427	585,383	567,208	574,822	466,013
$Y_{t-1} - Y_t$		-17,319	-21,758	-1,776	-28,789	-27,044	-18,175	7,614	-108,809
모형 C	634,475	650,450	635,982	627,679	627,448	603,799	592,558	563,691	543,820
$Y_{t-1} - Y_t$		15,975	-14,468	-8,303	-231	-23,649	-11,241	-28,867	-19,871
교육-모형C	12,956	14,300	7,010	13,537	-15,021	-18,416	-25,350	11,131	-77,807
연도	2018	2019	2020	2021	2022	2023	2024	2025	2026
교육개발원	451,563	427,950	406,817	407,857	408,523				
$Y_{t-1} - Y_t$		-14,450	-23,613	-21,133	1,040	666			
모형 C	547,225	481,118	427,879	425,098	408,797	376,126	387,109	426,400	403,892
$Y_{t-1} - Y_t$		3,405	-66,107	-53,239	-2,781	-16,301	-32,671	10,983	39,291
교육-모형C		-95,662	-53,168	-21,062	-17,241	-274			

※  $Y_{t-1} - Y_t = (t - 1)$ 연도 고3학생수 -  $t$ 연도 학생수이고,

교육-모형C= 교육개발원 고3학생수 예측결과 - 모형 C의 고3학생수 예측 결과이다.

## 참고문헌

- 김연형 (1994). <시계열분석>, 자유아카데미, 서울.
- 김종태, 서효민, 이인락 (2009). 2026년까지 대구광역시와 경상북도 지역의 고등학교 3학년 학생수에 대한 예측과 대학입학정원수와의 비교. <한국데이터정보과학회지>, **20**, 159-169.
- 김종태 (2009). 홀트-윈터스 가법모형에 의한 전국 학생수 예측. <한국데이터정보과학회지>, **20**, 685-694.
- 조찬혁 (2008). 동해항 시멘트 물동량의 추정에 관한 연구. <물류학회지>, **18**, 33-53.
- 통계청 (2006). <장래인구특별추계결과>, 통계정보시스템, 대전.
- 한국교육개발원 (2005). <교육통계 예측 결과>, 교육통계서비스, 서울.
- 한국교육개발원 (2006). <교육통계 예측 결과>, 교육통계서비스, 서울.
- 한국교육개발원 (2007). <교육통계 예측 결과>, 교육통계서비스, 서울.
- 한국교육개발원 (1982-2008). <교육통계연보>, 교육통계서비스, 서울.
- BakiBillah, M., King, M. L., Snyder, R. D. and Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, **22**, 239-247.
- Kim, J. T. (2005a). The forecasting about the numbers of the third graders in a high-school until 2022 Year in Daegu. *Journal of the Korean Data and Information Science Society*, **16**, 933-942.
- Kim, J. T. (2005b). The forecasting for the numbers of a high-school graduate and the number limit of matriculation in Kyungbook. *Journal of the Korean Data and Information Science Society*, **16**, 969-977.

# The methods of forecasting for the number of student based on promotion proportion

Jongtae Kim<sup>1</sup>

<sup>1</sup> Department of Computing & Statistics, Daegu University

Received 13 July 2009, revised 16 September 2009, accepted 21 September 2009

## Abstract

The purpose of this paper is to suggest the methods of forecasting for the number of the elementary, middle and high-school student based on the proportion of promotion until 2026 year. The suggested methods are the proportion of promotion, moving average, Holt-Winters model, SARIMA, regression fit. As the result, the abilities of forecasting by the method of moving average are better than those of other methods.

*Keywords:* Holt-Winters model, moving average, regression, time series.

---

<sup>1</sup> Professor, Department Computing & Statistics, Daegu University, Kyongsan 712-714, Korea.  
E-mail: jtkim@daegu.ac.kr