인구추계 데이터의 이상점과 통계적 분석

김종태*, 서효민**
(Jong-tae Kim , Hyo-Min Seo)

요 약 The purpose of this paper is to suggest the problems of basic population data(1960-2005) and the data(2006-2050) of population projections reported by Korean National Statistical Office in November 2006. The errors on the basic population data can be easily checked by using the graphical analysis and the method of linear regression analysis. It is necessary to revise the population projections reported by Korean National Statistical Office.

핵심주제어 : 결정계수, 선형회귀, 인구통계, 장래인구추계, 통계정보시스템

Key Words : Linear regression, Population projection, Statistical information system

1. 연구의 배경과 서론

본 연구는 Yoon과 Kim(2007)의 연구를 이용하여, 1960년 인구추계 통계에서, 0세에서 40세까지 각 연령별 인구수가 2005년에 45세에서 80세가 되기까지 46년간의 각 연령별 인구수에 따른 변화를 분석한 것이다. 그리고 2000년에서 1961년 사이의 0세 인구수가 해가 더해감에 따라서 2005년 5세에서 44세가 되기까지의 각 연령별 인구에 따른 변화를 분석한다. 2005년을 기점으로 정한 이유는 통계청(2006)의 '전국인구추계'통계가 1960년에서 2005년까지는 우리나라 '확정인구'로, 2006년부터 2050년까지는 '장래인구'로 추계하였기 때문이다.

통계청의 '공식통계'인 '확정인구'와 '장래인구'는 매우 중요한 통계자료이다. 이 통계 자료를 사용하여 인구와 관련된 분야-사회, 정치, 군사, 행정, 경제, 선거, 산업, 보건, 복지, 교육 문화 등-에서 정책 연구와 기획에 중요한 자료 역할을 하기 때문이다. 예로서 장래인구추계 및 고령화의 예측, 국민연금을 비롯한 각종 연금들의 고갈, 행정구역의 인구의 변동대책, 군입대자, 초등, 중등학교 의무교육, 지역별 선거인명부, 교육, 행정, 산업 문화, 예술, 체육 정책들과 인프라 등이 있다.(전광희(2006), Kim(2007)).

그러므로 '확정인구'와 '장래인구' 통계에 문제점과 이상점이 있다면, '인구추계' 통계를 사용한 정책에 오류가 발생할 확률이 높아지고, 심한 경우에는 정책에 대한 신뢰성을 잃어버리거나 잘못된 결과들을 얻게 된다. 이러한 이유로 '인구추계' 통계가 통계학적으로 신뢰할 만한 데이터인지를 검정할 필요가 있다.

기존의 인구통계에 대한 연구문헌으로는 다음과 같다. 인구성장모형에 대한 기존연구에는 선형모형에서는 Mansfield-Blackman모형, Linear Gompertz모형, Weibull모형, Nonsysmmetic Reponding Logistic (NSRL) 모형, Harvey모형 등이 있고, 비선형모형에는

* 경북 경산시 진량면 내리동 대구대학교 전산통계학과 교수 E-mail: jtkim@daegu.ac.kr
** 경북 경산시 진량면 내리동 대구대학교 통계학과 석사과정

지수기하모형, 수정지수곡선모형, Logistic 모형, Probit 모형, Gompertz 모형 등이 있다.(최종후외 3인(2000), Young(1993)). 인구추계의 분석모형으로 시계열분석방법인 추세연장법, 평활법, ARIMA모형, 백터자기회귀모형 등이 있다. 주로 사용하는 추세연장법은 사용대상에 따라 인구추세, 비율추세, 차이추세, 밀도추세로 세분화 된다.(오재석(2003)). Yoon과 Kim(2007)은 장래인구추계에 대하여 이상점, 특이성들을 조사하였고, 데이터의 이상점들이 2006년부터 2050년까지의 장래인구추계에서도 나타남을 발견하였다.

연령의 변화에 따른 각 연령별 인구분석은 그래픽적인 통계분석이나 선형회귀모형에 기초한 기울기, 결정계수, 분산 등을 이용하여도 연령이동에 따른 인구수의 증가에 대한 오류나 문제점들을 쉽고 충분히 발견하게 해 준다. 이러 분석방법은 기존인구데이터의 문제점과 이상값을 쉽게 찾게 뿐만 아니라 인구추계도 가능하게 한다.

2절에서는 통계청의 장래인구추계 방법을 설명하고, 확정된 기존인구에 대한 문제점들을 그래픽 분석을 이용하여 분석하였다. 3절에서는 선형회귀모형의 기울기들의 변화와 결정계수들을 사용하여 확정된 기존인구와 장래인구에 대한 분포를 비교 분석한다. 마지막으로 결론을 소개한다.

2. 추계인구와 그래픽 분석

통계청은 5년마다 인구추계를 실시한다. 통계청은 2006년 12월에 장,단기 국가발전의 계획수립과 향후 인구관련 각종 경제, 사회지표를 제공하기 위한 기초자료 및 학술자료를 제공하기 위한 목적으로, 2005년 인구주택총조사 결과를 기초로 인구변동요인(출생, 사망, 국제이동)별 실적자료 추이를 반영하여 2006년부터 2050년까지의 장래인구추계 결과를 추정한 '장래인구추계' 보고서를 발표하였다. 1960년부터 2005년까지의 '확정인구'와, 2006년부터 2050년까지 '장래추계인구'를 산출하여 통계정보시스템(KOSIS)에서 제공하고 있다.

'장래인구추계 2006.12'의 특징으로 법무부의 외국인 등록 체류자 및 불법 체류자 자료를 활용하여 정확한 인구기준을 작성하였고, 가정설정의 예측오차를 줄여 추계의 불확실성을 개선하고자, 출생은 로그감마모형, 사망은 Lee-Carter 및 Brass Logit 방법을 이용하여 작성하였다고 발표했다. 이 연구들을 기초로 특정 연도의 성 및 연령별 기존인구에 인구변동 요인인 출생, 사망, 국제이동에 대한 장래변동을 추정하여 조합하는 방법인 코호트요인법을 이용하여 장래인구추계를 하였다.

통계청(2006)의 '확정인구' 통계는 다음과 같은 절차와 방법으로 결정되었다. 2000년의 과거 확정인구수를 기준으로, 2000-2005년간의 출생, 사망, 국제이동에 관한 실적자료를 감안하여, 2005년 총인구를 보정하는 인구균형방정식(Demographic Balancing Equation)에 총인구를 보정함으로 기존인구를 작성한다. 이러한 확정된 기존인구수를 결정할 때, 연차별 인구총조사 출생코호트를 비교하고, 사후조사에 의한 오차율을 검토하고, 인구동태 및 주민등록인구 등과 비교한 후 인구변동요인에 의한 총인구를 확정하고 성 및 연령별 인구를 보정한다. 이런 절차를 걸친 후에 가중이동평균을 이용한 연령평활을 하고 총 조사시점인 (11월1일) 인구를 연양인구 (7월1일) 기준으로 환산하여 2001-2004년 추계인구를 실적자료로 대체한다.

위의 관점에서 본다면 통계청의 장래인구추계는 매우 신뢰성 있는 통계 자료임이 분명하다. 실제로 통계청이 집계하고 발표한 '인구추계' 통계 외에 '인구 총 조사'나 '주민등록인구' 통계를 제외한다면 기존에 신뢰할만한 인구통계 데이터를 구하기가 쉽지가 않다.

다음과 같은 그래픽 분석에 의해서 통계청의 '인구추계'에서 확정인구 통계에 대한 검정을 하여 보자. Yoon과 Kim(2007)은 1960년부터 2005년까지의 확정인구 통계에 대한 검정을 하기 위하여 아래와 같은 일반적인 가정들을 수립했다. (단, 국내거주 외국인수는 인구통계집계에 포함하지 않는다.)

[가정1] (Y+1)년도의 (G+1)세 인구수는 Y년도의 G세 인구수는 보다 적어야 한다.

$$\frac{\text{(Y+1)년도의 (G+1)세 인구수}}{\text{Y년도의 G세 인구수}} < 1.$$

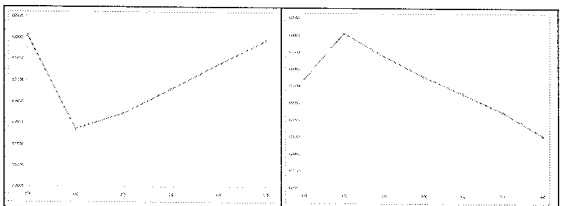
이는 사망으로 인하여, Y년도의 G세 인구가 (Y+1)년도의 (G+1)세가 되었을 때, 일반적으로 인구가 감소하고, 해외 이민자 수가 국적 취득자 보다 많기 때문이다.

[가정2] (Y+1)년도의 (G+1)세 인구수는 Y년도의 G세 인구수와 비교할 때, 특별한 이유가 없는 한 인구수의 변동이 매우 적거나 클 수가 없다.

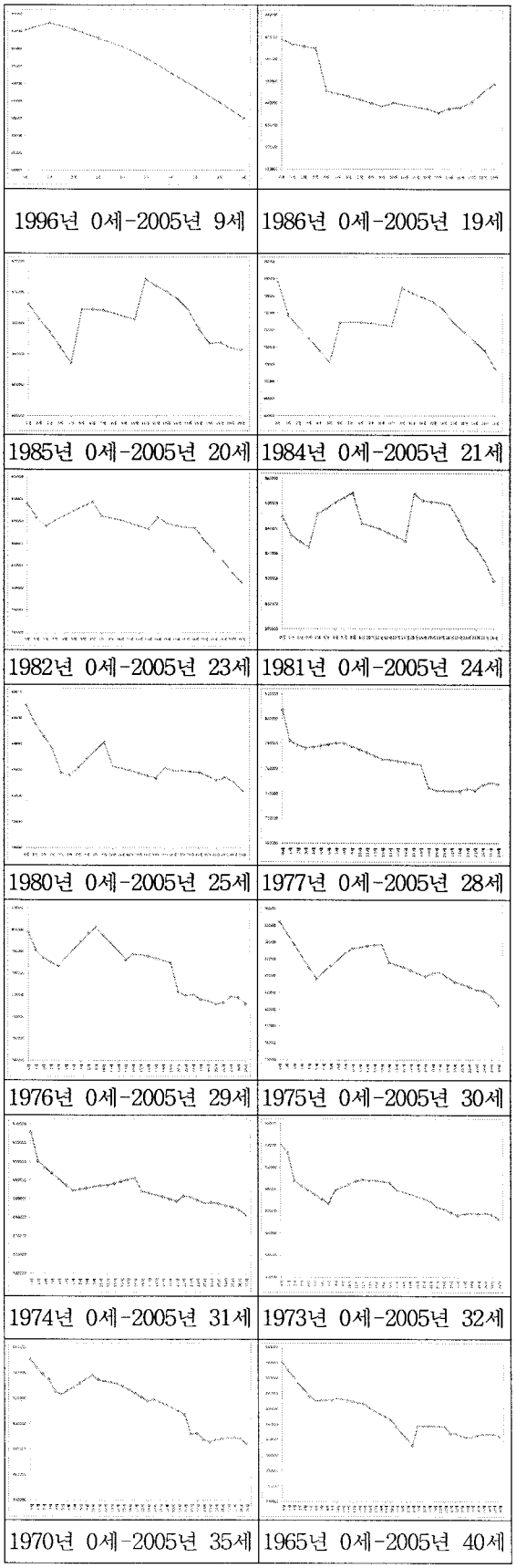
$$\left|\frac{\text{(Y+1)년도의 (G+1)세 인구수}}{\text{Y년도의 G세 인구수}}\right| < C.$$

여기서 C 값은 이상점 판정을 위한 상수값이다. 특별히 국가적 재앙이나 전쟁, 전염병, 대규모 해외이주 등의 대규모의 인구 손실이 있지 않는 한은 $(Y+1)$년도의 $(G+1)$세 인구수는 Y년도의 G세 인구수와 비교할 때, 그 변동이 C값 보다 크면 인구집계 통계에 오류가 발생한 것으로 본다.

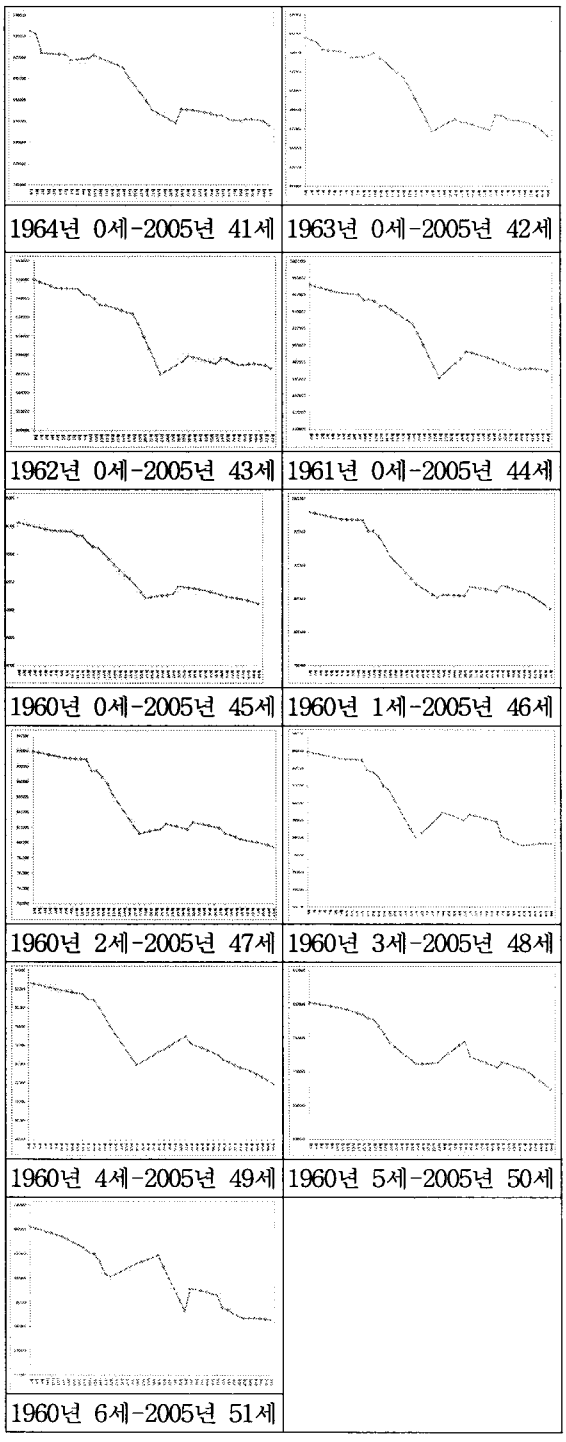
아래의 [그림 1]은 1960년에서 2000년 사이의 0세 인구가 2005년에 5세에서 45세가 되는 연령층들에 대한 인구수의 변동을 조사한 것 중 일부분이다. 또한 1960년 1세에서 6세의 연령들이 2005년 46세에서 51세가 되는 연령층에 대한 인구수의 변동을 조사한 것 중 일부분이 포함되었다. [그림1]에서 보여 준 연령층에 따른 인구수 변동에 대한 그래프들은 위의 [가정1]을 모두 위반하고 있으며 심지어 [가정2]를 위반하는 경우가 있음을 보여 준다.



2000년 0세-2005년 5세 | 1999년 0세-2005년 6세



1996년 0세-2005년 9세 | 1986년 0세-2005년 19세
1985년 0세-2005년 20세 | 1984년 0세-2005년 21세
1982년 0세-2005년 23세 | 1981년 0세-2005년 24세
1980년 0세-2005년 25세 | 1977년 0세-2005년 28세
1976년 0세-2005년 29세 | 1975년 0세-2005년 30세
1974년 0세-2005년 31세 | 1973년 0세-2005년 32세
1970년 0세-2005년 35세 | 1965년 0세-2005년 40세



[그림1] 각 연령별 연령증가에 대한 인구수 그래프

통계청의 '추계인구'에서의 확정 인구 통계에 대하여 [그림1]에서 다음과 같은 의문점을 던지지 않을 수 없다.

(1) 어느 특정한 연령층의 특정시기 동안에 나이가 듦에 따라서 그 인구수가 왜 꾸준히 증가하는가? 그 증가분에 대한 인구는 어디서 생겨난 것인가?

(2) 어느 특정한 연령층의 특정시기 동안에 나이가 듦에 따라서 그 인구수가 왜 갑자기 크게 감소했다가 다시 크게 증가하는가? 심지어 이 현상을 반복하는 특정 연령층들도 있음을 보여 준다.

(3) 어떤 연령층은 연령이 충분히 들었음에도 출생아 수 보다 더 많은 인구수를 나타낸다. 태어나지 안했던 인구는 어디서 생긴 것인가? 0세 인구수에 대한 집계오류일가?

위의 (1), (2), (3)의 경우를 고려할 때, 통계청의 '인구추계' 통계의 확정 인구추계는 신뢰성이 있는 통계인지에 대한 의문을 가지지 않을 수 없다. 이러한 기존 확정인구의 오류는 '장래추계인구'를 예측할 때, 나이가 듦에 따라서 꾸준히 지속적으로 인구가 증가하는 현상을 보이는 연령층도 있다. 이러한 문제점이 발생하는 이유는 추측하건데 다음과 같다.

1) 추계인구는 5년마다 확정을 하는데 앞서의 기존인구의 보정작업 잘못된 경우,

2) 통계집계에서 기초자료가 잘못된 경우.

3) 통계적 전문적인 지식 없는 경우 등을 추측할 수 있다.

3. 선형회귀모형에 따른 장래추계인구 분석

2절에 있는 [가정1]과 [가정2]의 조건 하에서, 연령의 변화에 따른 각 연령별 인구수의 변화에 대한 모형으로 다음의 선형회귀모형을 고려한다.

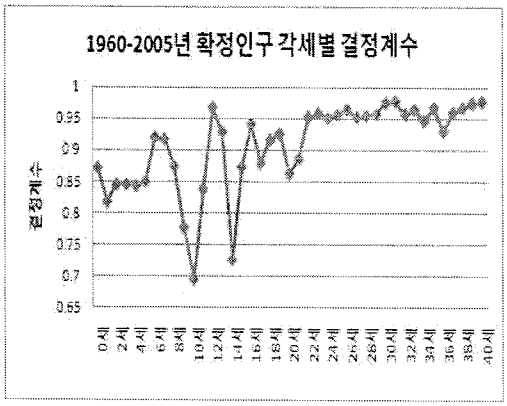
$$y_{ik} = \beta_{0i} + \beta_{1i} x_{ik} + \epsilon_{ik}, \quad i = 0, \ldots, m, \quad k = 0, \ldots, n. \qquad (3.1)$$

여기서 i는 기준연도의 0세에서 m세까지의 연령을 나타내고, k는 0을 기준 되는 해로 정하고, 1년씩 증가하여 n까지 연도수를 나타낸다. x_{ik}는 i세의 기준연도의 연령층들이 한 해씩 증가하는 연도이고, y_{ik}는 i세의 연령층들이 나이가 듦에 따라 x_{ik}가 되는 해의 인구수를 의미한다. ϵ_{ik}는 i, j 번째 관측의 확률오차로서 평균이 0이고 분산 σ^2인 정규분포를 따른다고 가정한다.

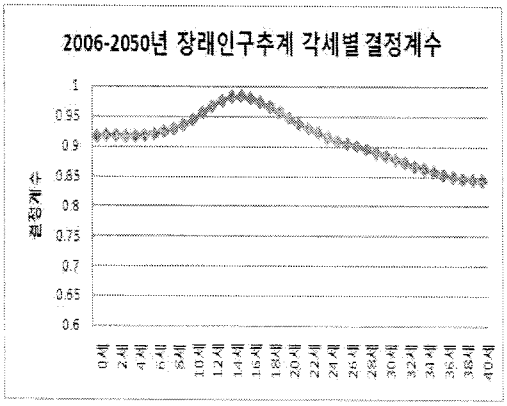
모형(3.1)의 적합성을 조사하기 위해 연령의 변화에 따른 각 연령별 인구수의 변화에 대한 선형강도의 관계를 나타내는 결정계수 R^2을 조사한다.

$$R_i^2 = \sum_{k=0}^{n}(x_{ik}-\overline{x_i})(y_{ik}-\overline{y_i}) / \left\{ \sum_{k=0}^{n}(x_{ik}-\overline{x_i})^2 \sum_{k=0}^{n}(y_{ik}-\overline{y_i})^2 \right\}, \quad i=1,\ldots,n \tag{3.2}$$

[그림2(a)]는 1960년 0세에서 40세까지의 각 연령층의 인구가 2005년에 45세에서 85세가 되기까지 46년 동안의 각 연령별 인구수의 변화에 대한 선형회귀모형에서의 결정계수 R^2들에 대한 그림이고, [그림2(b)]는 2006년 0세에서 40세까지의 각 연령층의 인구가 2060년에 44세에서 84세가 되기까지 45년 동안의 각 연령별 인구수의 변화에 대한 선형회귀모형에서의 결정계수 R^2들에 대한 그림이다.



[그림2(a)] 연령별 연령증가 결정계수



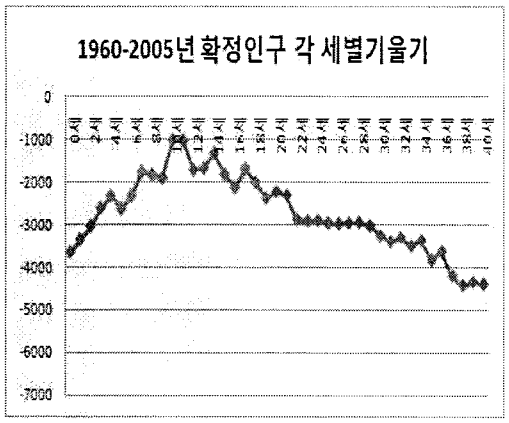
[그림2(b)] 연령별 연령증가 결정계수

기존 확정된 인구에 대한 [그림2(a)]에서 결정계수 값은 9, 10, 14세에서 0.8 이하이고, 0-5세, 8, 11, 15, 20, 21세는 0.8~0.9 사이에 있고 나머지 연령층은 0.9 이상이다. 이는 거의 대부분의 연령층들이 (3.1)에 주어진 선형모형에 잘 적합함을 보여준다. 특히 22세 이후의 연령층에 대한 결정계수의 값들은 대부분 0.9 이상으로 연령이 높아질수록 거의 선형에 가까운 분포를 나타낸다. 그러나 9, 10, 14세의 연령층의 인구변동에 결정계수의 값은 0.8이하이고 약한 선형강도를 나타낸다.

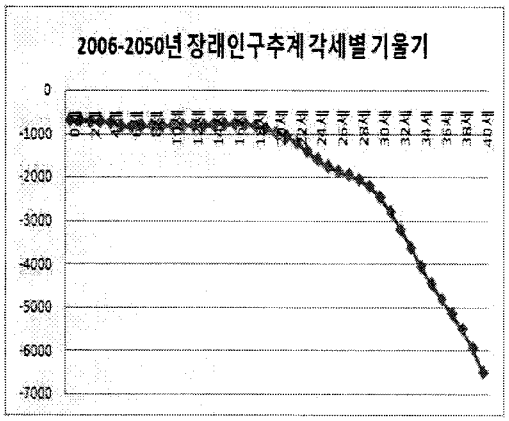
장래추계인구에 대한 [그림2(b)]와 [그림2(a)]를 비교하면, [그림2(b)]는 결정계수의 변동이 매우 스무드한 반면에 [그림2(a)]는 6-22세까지 연령층들의 결정계수의 변동이 매우 크게 나타나고 있다. 이러한 결정계수의 갑작스러운 변동은 확정된 인구수의 추계에 대하여 어떤 문제점이 있다는 것을 암시하고 있다.

그리고 [그림2(b)]에서 15세 이후에서 결정계수의 값이 점차적으로 떨어지는 것이 흥미롭다. 기존의 확정인구에서는 1960년 22세 이후의 연령층들이 46년 지나는 동안 2005년까지 거의 선형 모형을 띄는 동안에, 2006년의 26세 연령층들은 44년이 지나 2050년이 되는 해까지 처음에는 조금씩 인구가 줄어들다가 뒤에는 급격히 인구가 줄어드는 현상으로 장래추계를 하였다. 과거 45년과 향후 44년의 결정계수의 분포가 매우 다르게 나타나고 있음을 알 수 있다.

이러한 차이는 선형회귀모형(3.1)의 β_{ik} 들을 추정한 기울기 값들의 변화에서도 나타난다. 아래의 [그림3(a)]는 1960년 0세에서 40세까지의 각 연령층의 인구가 2005년에 45세에서 85세가 되기까지 46년 동안의 각 연령별 인구수의 변화에 대한 선형회귀모형에서의 기울기 값들에 대한 그림이고, [그림3(b)]는 2006년 0세에서 40세까지의 각 연령층의 인구가 2060년에 44세에서 84세가 되기까지 45년 동안의 각 연령층별 인구수 변화에 대한 선형회귀모형에서의 기울기 값들에 대한 그림이다.



[그림3(a)] 연령별 연령증가 기울기



[그림3(b)] 연령별 연령증가 그래프

확정된 기존추계인구와 장래인구추계의 차이점을 살펴보면 다음과 같다.

[그림3(a)]는 1960년 0-10세 사이의 연령층별 인구감소에 따른 선형회귀모형의 기울기 값들은 가파르게 커지고 있음을 보인다. 10-22세의 연령층 사이의 기울기 값들은 감소와 증가를 반복하면서, 결국 작아지는 경향을 나타내고 있다. 22세 이후 연령층 기울기 값들은 서서히 작아지는 경향을 나타낸다.

[그림3(b)]는 2006년을 기점으로 향후 45년간의 선형회귀모형의 기울기 값들의 변화를 나타낸 것으로, 과거 46년간의 선형회귀모형의 기울기 값들의 변화인 [그림3(a)]와는 다르게 나타낸다. [그림3(b)]에서, 0-19세의 연령층의 기울기의 변화는 거의 없지만, 20-40세 사이의 연령층의 기울기 값들은 매우 심하게 작아지는 현상을 보인다.

과거 46년간의 0-40세까지의 연령층별 인구감소에 따른 기울기의 변화는 적어도 -4,000보다 크다. 그러나 향후 45년 동안에는 33세 연령층을 기점으로 기울기 값은 -4,000 이하로 작아지면서, 40세 연령층에서는 -6,500의 심한 인구 감소 현상의 변동을 나타낸다. 의학의 발달로 인간의 수명이 늘어나는 현상을 고려할 때, 33세 연령층 이후의 인구감소율이 과거 46년보다 향후 45년간이 더 심하게 작아지는 것은 장래인구추계이론은 모순이 있어 보인다.

4. 결론

각 연령층별 전국인구는 외국에서 수입되지 않는 한은 갑자기 늘어나지 않는다. 그러나 확정된 기존 인구통계(1960-2005)에서 이러한 현상이 많이 보인다. 또한 천재지변이나, 전쟁 등으로 인한 많은 사람들의 사망이나 국적포기가 일어나지 않는 이상 각 연령층별 인구가 갑자기 이상점 이하로 감소할 수 없다. 그러나 이러한 현상도 확정된 기존 인구통계에서 자주 발견된다.

통계적 추론이란 과거 혹은 기존의 데이터를 가지고, 미래의 예측을 한다. 그러나 선형회귀모형에서의 결정계수와 기울기의 분포를 고려할 때, 과거 46년간의 인구자료와 향후 45년간의 예측인구 사이에는 매우 다른 형태의 분포들을 보인다.

이러한 확정된 기존인구추계와 장래인구추계에 대해서는 통계적으로 정확하고 신뢰성이 있는 분석이 요구되어진다. 서론에서 언급한 대로 통계청의 '공식통계'인 '확정인구'와 '장래인구'는 매우 중요한 통계자료이고, 이 통계자료를 이용하여 인구와 관련된 사회, 정치, 군사, 행정, 경제, 선거, 산업, 보건, 복지, 교육 문화 등의 분야에서 정책 연구나 기획에서 중요한 자료 역할을 하기 때문이다.

결론적으로 "과연 우리나라 각 연령별 인구통계는 어느 것이 가장 정확한 인구통계일까?"하는 것에 의문을 던지지 않을 수 없다. 고대로부터 인구통계는 한 나라의 가장 기본이 되고 중요한 데이터임에는 틀림없는 사실이므로 정확하고 세밀한 인구통계 데이터의 보정이 필요하다.

참 고 문 헌

[1] 오재석 (2003). 통계적 인구추정모형의 정립과 검증에 관한 연구, 조선대학교, 석사과정학위논문.

[2] 전광희 (2006). 한국의 혼인력과 출산력 추이와 전망. 대한통계협회, 제32권 1, 2호 통합본, p60-90.

[3] 최종후, 최봉호, 양우성, 김유진(2000). 성장곡선모형에 의한 인구예측시스템, 한국인구학회, 제23권, 제1호, 197-215.

[4] 통계청 (2006). '장래인구추계' 보고서, 2006.11.

[5] Kim, J.T. (2007). The Errors of Forecast Educational Statistics on Koean National Center for Educational Statistics & Information. Journal of the Korean Data and Information Science Society, Vol. 18, 141-148.

[6] Yoon, Y.H. and Kim, J.T. (2007). The Errors of Population Projections for Korea on Korean Information Statistical System. Journal of the Korean Data and Information Science Society, Vol. 18, 419-427.

[7] Young P. (1993). Technological Growth Curve: A Competition of Forecasting Models, Technological Forecasting and Social Chance, Vol.44, 375-389.