

비선형 회귀모형을 이용한 학년별 학생수 추계[†]

윤용화¹ · 김종태²

^{1,2}대구대학교 전산통계학과

접수 2011년 12월 2일, 수정 2011년 12월 21일, 게재확정 2011년 12월 28일

요약

본 연구는 코호트 조성법에 의해 구성된 진학률들을 사용한 비선형 회귀모형을 이용하여 장래 초 등과 중등, 고등학교의 학년별 학생수를 추계 하는데 목적이 있다. 이러한 진학률들의 모형을 분석하기 위하여 경향-외삽법 중 하나인 비선형 회귀모형의 로그모형과 거듭제곱 모형을 이용하였다. 그 결과 로그모형에 의한 예측이 거듭제곱모형에 의한 예측보다 조금 더 신뢰할 수 있고, 학생수도 적게 예측됨을 알 수 있었다.

주요용어: 거듭제곱 모형, 로그 모형, 비선형 회귀모형.

1. 서론

본 연구는 인구추계를 위한 경향-외삽법 (trend-extrapolation methods)의 한 종류인 비선형 회귀모형을 이용하여 전국 학년별 장래 학생수 추계를 하였다. 장래 학생수 예측의 방법으로써 비선형 회귀모형인 로그모형 (log model)과 거듭제곱모형 (involution model)을 사용하였다.

장래인구추계는 대부분의 경우에는 5세 단위로 추계하여 왔다. 추계 방법으로는 인구의 구성요소인 출생, 사망, 인구이동에 관한 정보를 이용하여, 각 연령집단별 (코호트) 인구변화의 요인이 계속 유지된다는 가정 하에 장래인구를 추계하는 코호트-조성법이 사용되고 있다 (이상림과 조영태, 2005). 그러나 출생, 사망, 인구이동의 정보를 얻는 것은 쉽지 않은 통계 작업이다. 또한 코호트-조성법에 의해 기초지방자치단체의 연령별 장래인구추계를 위해 필요한 인구동태 데이터들을 찾는 데는 많은 어려움이 있다.

경향-외삽법의 대표적인 방법들로서 미래인구를 예측하는 방법으로 선형모형, 기하모형, 지수모형, 로지스틱모형, 다항곡선모형, ARIMA 시계열모형 등을 들고 있다. 20세기 중반까지 사용되었던 경향-외삽법은 전체 인구수 추계만 가능할 뿐 성별 연령별 인구구조의 추계가 불가능하다는 치명적인 약점으로 인해 코호트-조성법으로 대체되었다. 이상림과 조영태 (2005)는 코호트별 인구변화율 (cohort-change ratio)이 일정하게 지속된다는 가정 하에서, Hamilton과 Perry (1962)의 기법을 이용하여 사망과 인구 이동에 대한 자료 없이 기존의 센서스를 바탕으로 하여 기초자치단체의 장래인구를 추계하였다.

Meade (1988)는 인구수를 추정하기 위한 방법으로 수정된 로지스틱모형을 제시했고, Raeside (1988)는 인구수의 경향을 찾기 위한 모형으로 로지스틱 모형을 이용하였다. 김종태 (2009a, 2009b)에서는 대구와 경북지역을 중심으로 고3학생수를 예측하고, 대구와 경북 지역의 대학 신입생 수와 비교 분

[†] 이 논문은 2009년도 대구대학교 학술연구비 지원에 의한 연구임.

¹ (712-714) 경상북도 경산시 진량면 내리동15, 대구대학교 전산통계학과, 교수.

² 교신저자: (712-714) 경상북도 경산시 진량면 내리동15, 대구대학교 전산통계학과, 교수.

E-mail: jtkim@daegu.ac.kr

석하였다. 송필준과 김종태 (2010)는 로지스틱모형과 비레이동평균모형에 의한 학년별 학생수 추계 방법을 소개하고, 비교분석 하였다.

본 연구에서는 이상립과 조영태 (2005)와는 달리 전국 코호트별 인구변화율과 코호트별 학년진학률을 비선형 모형인 로그모형과 거듭제곱모형을 사용하여 학년별 장래학생수를 추계하였다. 결론적으로 로그모형에 의한 결과가 거듭제곱모형에 의한 결과보다 장래학생수를 더 적게 예측하였고, 조금 더 신뢰할 수 있다는 결과를 얻었다.

2절에서는 진학률 예측을 위한 비선형 회귀모형을 소개하고, 3절에서는 비선형 회귀모형에 의한 장래 학생수의 추계와 그 결과들을 비교 분석하였다.

2. 진학률예측을 위한 비선형 회귀모형

기존의 학년 (연령)별 진학률 코호트를 살펴보면 선형모형을 사용하는 것보다는 선형식이 아닌 비선형식 (nonlinear equation)인 로그 (log) 모형이나 거듭제곱 모형을 적용할 때, 진학률 코호트에 잘 적용됨을 실험적 연구를 이용하여 알 수 있다.

주민등록 0세 인구수 추정에 있어서 결정계수 R^2 의 값을 비교해 볼 때, 로그 모형의 결정계수의 값은 82.09%이고, 거듭제곱의 결정계수의 값은 81.21%, 선형모형에 의한 결정계수의 값은 74.93%이다. 그러나 선형모형을 장래인구추정에 사용한다면 15년 후에는 주민등록 0세 인구수가 0명 이하로 추정되는 적절하지 못한 결과가 발생한다. 로그 모형과 거듭제곱모형을 비교할 때, 직관적으로 거듭제곱모형이 로그모형 보다 인구수를 과도하게 추정한다. 비선형 회귀 모형에 대한 설명은 일반적인 회귀분석 교과서에서 잘 설명되어 있다. 로그 모형인 경우는 다음과 같다.

$$y = \alpha_0 + \alpha_1 \ln(x) + \epsilon \quad (2.1)$$

이 되며, $\alpha_0 = \beta_0$, $\alpha_1 = \beta_1$, $\ln(x) = x'$ 으로 두면,

$$y = \beta_0 + \beta_1 x' + \epsilon$$

으로 선형 회귀직선이 된다. 그러면 α_0 와 α_1 의 추정치 $\hat{\alpha}_0$ 와 $\hat{\alpha}_1$ 과 β_0 와 β_1 의 추정치 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 에 대하여,

$$\hat{\alpha}_0 = \bar{y} - \hat{\beta}_1 \bar{x}', \quad \hat{\alpha}_1 = \frac{\sum_i (x'_i - \bar{x}') (y_i - \bar{y})}{\sum_i (x'_i - \bar{x}')^2} \quad (2.2)$$

가 되고 따라서 식 (2.1)의 추정값 \hat{y} 는 다음과 같다.

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \ln(x). \quad (2.3)$$

거듭제곱모형은 다음 식 (2.4)과 같다.

$$y = \alpha_0 x^{\alpha_1} \epsilon \quad (2.4)$$

식 (2.4)을 로그변환을 시키면,

$$\log(y) = \log(\alpha_0) + \alpha_1 \log(x) + \log(\epsilon) \quad (2.5)$$

로 되고, $\log(y) = y'$, $\alpha_1 = \beta_1$, $\log(\alpha_0) = \beta_0$, $\log(x) = x'$, $\log(\epsilon) = \epsilon'$ 으로 두면, 회귀직선은 다음과 같다.

$$y' = \beta_0 + \beta_1 x' + \epsilon'.$$

α_0 와 α_1 의 추정치 $\hat{\alpha}_0$ 와 $\hat{\alpha}_1$ 과 β_0 와 β_1 의 추정치 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 에 대하여,

$$\hat{\alpha}_0 = \text{antilog}(\hat{\beta}_0), \hat{\alpha}_1 = \hat{\beta}_1$$

으로 구해지며, 식 (2.4)의 추정된 회귀직선은 다음과 같다.

$$\hat{y} = \hat{\alpha}_0 x^{\hat{\alpha}_1}.$$

로그모형과 거듭제곱모형을 이용한 2011년부터 2046년까지의 주민등록 0세인구수의 추정의 결과는 표 2.1과 표 2.2에서 나타내었고 이 값들은 장래 학생수 추계를 위한 초기값으로 사용된다.

표 2.1과 표 2.2를 살펴보면 로그모형과 거듭제곱모형을 이용한 주민등록 0세인구수 추정의 결과는 거듭제곱모형의 결과가 2011년에는 약 4,800명 정도로 로그모형보다 많았다. 그리고 추정인구수의 차이는 2046년에는 약 35,500명 정도로 차이가 벌어진다. 결정계수를 가지고 비교해 볼 때도, 거듭제곱모형의 결과 보다는 로그모형에 의한 장래 인구수 추정 결과가 보다 적절한 것으로 나타난다.

0세에서 5세까지는 인구변화율을 구하고, 초등1학년부터 고등3학년까지는 진학률을 구한다. 이들 인구변화율과 진학률을 이용하여 로그모형과 거듭제곱모형을 이용한 장래 0세에서 5세까지의 인구변화율과 초등1학년부터 고등3학년까지의 진학률들을 각각 추계한다.

표 2.1 로그모형을 이용한 주민등록 0세인구수 추정

2011	2012	2013	2014	2015	2016	2017	2018	2019
413,593	407,124	401,102	395,469	390,178	385,189	380,470	375,992	371,734
2020	2021	2022	2023	2024	2025	2026	2027	2028
367,674	363,794	360,079	356,516	353,093	349,798	346,624	343,561	340,602
2029	2030	2031	2032	2033	2034	2035	2036	2037
337,740	334,969	332,283	329,678	327,148	324,689	322,297	319,970	317,702
2038	2039	2040	2041	2042	2043	2044	2045	2046
315,493	313,337	311,234	309,180	307,174	305,212	303,294	301,417	299,579

표 2.2 거듭제곱모형을 이용한 주민등록 0세인구수 추정

2011	2012	2013	2014	2015	2016	2017	2018	2019
418,395	413,157	408,340	403,884	399,743	395,878	392,256	388,850	385,638
2020	2021	2022	2023	2024	2025	2026	2027	2028
382,600	379,720	376,983	374,376	371,888	369,509	367,232	365,048	362,950
2029	2030	2031	2032	2033	2034	2035	2036	2037
360,932	358,990	357,117	355,309	353,562	351,873	350,238	348,654	347,117
2038	2039	2040	2041	2042	2043	2044	2045	2046
345,627	344,179	342,772	341,403	340,072	338,775	337,511	336,280	335,078

추정된 장래 진학률들 각각에 대하여 표 2.1과 표 2.2의 결과를 초기값으로 사용하여 1세의 인구변화율에 곱해 줌으로써 주민등록 1세의 인구수를 추계하고, 추계된 1세의 인구수에 2세의 인구변화율에 곱해 줌으로써 주민등록 2세의 인구수를 추계하고, 이러한 과정을 반복하여 추계된 고등3학년 진학률에 추계된 고등2학년 학생수를 곱해주면 고등3학년 학생수가 추계된다.

3. 비선형 회귀모형에 의한 장래 학생수 추계

아래의 표 3.1와 표 3.2는 로그모형에 의하여 2011-2046년까지의 전국 초등, 중·고등 학생수를 각각 예측한 것이고, 표 3.3과 표 3.4은 거듭제곱모형에 의하여 2011 - 2046년까지의 전국 초등 중·고등 학생수를 각각 예측한 것이다.

표 3.1 로그모형에 의한 2011-2046년까지의 전국 초등 학생수 예측

전국	초1학년	초2학년	초3학년	초4학년	초5학년	초6학년
2000	703,112	720,388	705,241	635,104	636,435	619,711
2010	476,286	467,816	533,502	604,859	598,653	617,978
2013	392,071	427,203	444,897	472,529	462,451	527,006
2016	417,267	442,082	400,717	388,739	422,145	439,402
2019	379,600	385,675	395,249	413,511	436,711	395,710
2022	361,720	366,460	370,920	376,023	380,888	390,262
2025	346,821	350,632	354,125	358,178	361,828	366,200
2028	334,069	337,194	339,994	343,313	346,129	349,584
2031	322,935	325,534	327,814	330,593	332,802	335,606
2034	313,063	315,247	317,123	319,489	321,241	323,557
2037	304,202	306,050	307,606	309,647	311,041	312,983
2040	296,170	297,741	299,036	300,816	301,925	303,569
2043	288,829	290,168	291,247	292,813	293,689	295,094
2046	282,073	283,214	284,113	285,500	286,184	287,391

표 3.2 로그모형에 의한 2011-2046년까지의 전국 중·고등 학생수 예측

전국	중1학년	중2학년	중3학년	고1학년	고2학년	고3학년
2000	611,196	613,921	635,422	632,822	695,156	743,490
2010	646,048	655,012	673,738	663,317	649,524	649,515
2013	596,890	591,014	608,984	627,661	620,665	631,331
2016	466,393	456,781	519,726	579,833	560,068	570,936
2019	383,754	417,143	433,600	453,018	432,892	487,459
2022	408,269	431,690	390,690	372,715	395,350	406,829
2025	371,303	376,626	385,487	396,492	409,158	366,687
2028	353,723	357,878	361,866	360,567	356,984	361,908
2031	339,079	342,435	345,572	343,472	339,227	339,820
2034	326,547	329,325	331,862	329,231	324,600	324,596
2037	315,607	317,949	320,044	317,045	312,183	311,787
2040	305,910	307,913	309,670	306,408	301,409	300,745
2043	297,208	298,941	300,433	296,979	291,903	291,052
2046	289,322	290,834	292,115	288,517	283,405	282,420

표 3.3 거듭제곱모형에 의한 2011-2046년까지의 전국 초등 학생수 예측

전국	초1학년	초2학년	초3학년	초4학년	초5학년	초6학년
2000	703,112	720,388	705,241	635,104	636,435	619,711
2010	476,286	467,816	533,502	604,859	598,653	617,978
2013	434,278	472,017	490,260	472,527	462,457	527,009
2016	465,212	491,869	444,882	430,585	466,435	484,209
2019	431,702	436,466	441,366	461,024	485,901	439,327
2022	417,268	420,764	423,871	427,634	431,057	435,801
2025	405,552	408,183	410,391	413,183	415,453	418,481
2028	395,735	397,739	399,317	401,449	402,948	405,134
2031	387,317	388,846	389,958	391,617	392,566	394,167
2034	379,967	381,123	381,880	383,185	383,725	384,899
2037	373,460	374,316	374,791	375,823	376,047	376,898
2040	367,633	368,241	368,489	369,305	369,278	369,878
2043	362,364	362,764	362,826	363,467	363,237	363,636
2046	357,562	357,786	357,692	358,189	357,792	358,027

표 3.4 거듭제곱모형에 의한 2011-2046년까지의 전국 중·고등 학생수 예측

전국	중1학년	중2학년	중3학년	고1학년	고2학년	고3학년
2000	611,196	613,921	635,422	632,822	695,156	743,490
2010	646,048	655,012	673,738	663,317	649,524	649,515
2013	596,869	590,987	609,105	627,863	620,788	631,358
2016	466,375	456,764	519,841	580,015	560,166	571,079
2019	425,049	460,885	477,924	453,168	432,979	487,596
2022	455,162	480,290	433,858	412,983	436,917	448,446
2025	422,251	426,211	430,577	442,213	455,341	407,232
2028	408,029	410,895	413,636	410,214	404,094	404,271
2031	396,483	398,626	400,591	396,376	389,593	388,468
2034	386,810	388,443	389,879	385,142	377,977	376,309
2037	378,515	379,772	380,828	375,728	368,337	366,328
2040	371,273	372,244	373,018	367,655	360,129	357,898
2043	364,862	365,608	366,165	360,607	353,002	350,626
2046	359,120	359,687	360,074	354,367	346,721	344,247

표 3.1와 표 3.3을 비교해보면 로그모형에 의한 초등 학생수 예측이 거듭제곱모형에 의한 초등 학생수 예측보다도 다음과 같이 적게 나타난다.

표 3.5 로그모형과 거듭제곱모형에 의한 2046년의 전국 예측 초등 학생수 차이

초등1학년	초등2학년	초등3학년	초등4학년	초등5학년	초등6학년
-75,489	-74,572	-73,579	-72,689	-71,608	-70,636

표 3.2와 표 3.4을 비교해보면 로그모형에 의한 중·고등 학생수 예측이 거듭제곱모형에 의한 중·고등 학생수 예측보다도 다음과 같이 적게 나타난다.

표 3.6 로그모형과 거듭제곱모형에 의한 2046년의 전국 예측 중·고등 학생수 차이

중1학년	중2학년	중3학년	고1학년	고2학년	고3학년
-69,798	-68,853	-67,959	-65,850	-63,316	-61,827

위의 표 3.5과 표 3.6에서 로그모형과 거듭제곱모형에 의한 예측방법의 차이에 따라 2046년에는 초등1학년 학생수 차이 약 75,500명을 시작으로 고등3학년 학생수 차이 약 61,800명으로 학년이 올라갈수록 점차 차이가 적어짐을 보이고 있다. 이는 주민등록 0세 인구에 대한 초기값 추정이 거듭제곱모형이 로그모형보다 과추정 (over-estimate)되어 지고, 또한 해가 갈수록 줄어드는 출산이수의 감소에 원인이 있다.

어느 예측결과가 좀 더 신뢰할 수 있는가하는 문제를 해결하기 위해서, 로그 모형과 거듭제곱모형의 결정계수 R_1^2 , R_2^2 들에 대한 신뢰구간을 비교해 보았다. \bar{R}_1^2 , \bar{R}_2^2 는 각각 로그모형과 거듭제곱모형의 연령별 인구증가율 혹은 진학률에 대한 결정계수 R_1^2 , R_2^2 들의 각각의 평균이고, $s_{R_1^2}$, $s_{R_2^2}$ 은 결정계수 R_1^2 , R_2^2 들의 각각의 표준편차라 하자. 로그모형의 결정계수 R_1^2 의 95% 신뢰구간은 $\bar{R}_1^2 \pm 1.96 \times s_{R_1^2} / \sqrt{n}$ 은 (0.3226, 0.1365)이고, 거듭제곱모형의 결정계수 R_2^2 의 95% 신뢰구간 $\bar{R}_2^2 \pm 1.96 \times s_{R_2^2} / \sqrt{n}$ 은 (0.3236, 0.1364)으로 로그모형에 의한 결정계수가 평균값에서 조금 더 높게 나오고, 신뢰구간의 영역도 좁다는 것을 알 수 있다. 현재의 인구감소율에 기초한다면 로그모형의 예측결과가 더 타당한 것으로 보인다.

표 3.7은 2010년 인구(학생)수를 기준으로, 로그모형과 거듭제곱모형에 의한, 2046년의 전국 예측 인구(학생)수 감소율을 나타낸 것이다.

로그모형에 의한 감소율이 거둬제공모형에 의한 감소율보다 더 크게 나타나고, 2010년의 인구 (학생) 수를 기준으로 볼 때, 2046년의 고등3학년 학생수는 로그모형이 약 43.5%, 거둬제공모형이 53%로 각각 56.5%, 47%가 줄어드는 것으로 보인다.

표 3.7 2010년 인구 (학생)수 기준 대비 로그모형과 거둬제공모형에 의한 2046년의 전국 예측 인구 (학생)수 감소율

전국	로그모형	거둬제공	전국	로그모형	거둬제공
0세	70.6%	78.9%	초4학년	47.2%	59.2%
1세	67.5%	75.3%	초5학년	47.8%	59.8%
2세	64.2%	71.4%	초6학년	46.5%	57.9%
3세	71.2%	78.9%	중1학년	44.8%	55.6%
4세	73.8%	81.6%	중2학년	44.4%	54.9%
5세	68.2%	75.3%	중3학년	43.4%	53.4%
초1학년	59.2%	75.1%	고1학년	43.5%	53.4%
초2학년	60.5%	76.5%	고2학년	43.6%	53.4%
초3학년	53.3%	67.0%	고3학년	43.5%	53.0%

참고문헌

- 김종태 (2009a). 학년진학률에 따른 학생수 예측방법. <한국데이터정보과학회지>, **20**, 857-867.
- 김종태 (2009b). 주민등록 0세-6세 인구의 역추정과 기존 인구통계와의 출생아수 비교. <한국데이터정보과학회지>, **20**, 1145-1153.
- 송필준, 김종태 (2010). 로지스틱함수모형과 비례이동평균모형에 의한 학생 수 추계와 분석. <한국데이터정보과학회지>, **21**, 503-511.
- 이상림, 조영태 (2005). H-P 기법을 이용한 기초자치단체의 장래인구추계. <한국인구학>, **28**, 149-172.
- Hamilton, P. S. and Perry, F. S. (1962). A short method for projecting population by age from one decennial census to another. *Social Force*, **41**, 163-179.
- Meade, N. (1988). A method logistic model applied to human population. *Journal of Royal Statistical Society A*, **151**, 491-498.
- Raeseide, R. (1988). The use of sigmoids in modelling and forecasting human population. *Journal of Royal Statistical Society A*, **151**, 499-513.

Estimations of the student numbers by nonlinear regression model[†]

Yonghwa Yoon¹ · Jongtae Kim²

^{1,2}Department of Computing & Statistics, Daegu University

Received 2 December 2011, revised 21 December 2011, accepted 28 December 2011

Abstract

This paper introduces the projection methods by nonlinear regression model. To predict the student numbers, a log model and an involution model as the kind of a trend-extrapolation method are used. Empirical evidence shows that a projection by log model is better than by involution model with the confidence interval estimations for the coefficients of determination.

Keywords: Involution model, log model, nonlinear regression model.

[†] This research was supported by the Daegu University Research Grant 2009.

¹ Professor, Department Computing & Statistics, Daegu University, Kyungbook 712-714, Korea.

² Corresponding author: Professor, Department Computing & Statistics, Daegu University, Kyungbook 712-714, Korea. E-mail: jtkim@daegu.ac.kr