▨연구논문

비선형 성장곡선 모형의 분석 절차에 대한 연구

황정연

한국전자통신연구원 초고속서비스연구실

A Study on the Analysis Procedures of Nonlinear Growth Curve Models

Jung-Yeon Hwang

Electronics and Telecommunications Research Institute, ATM Service Section

Abstract

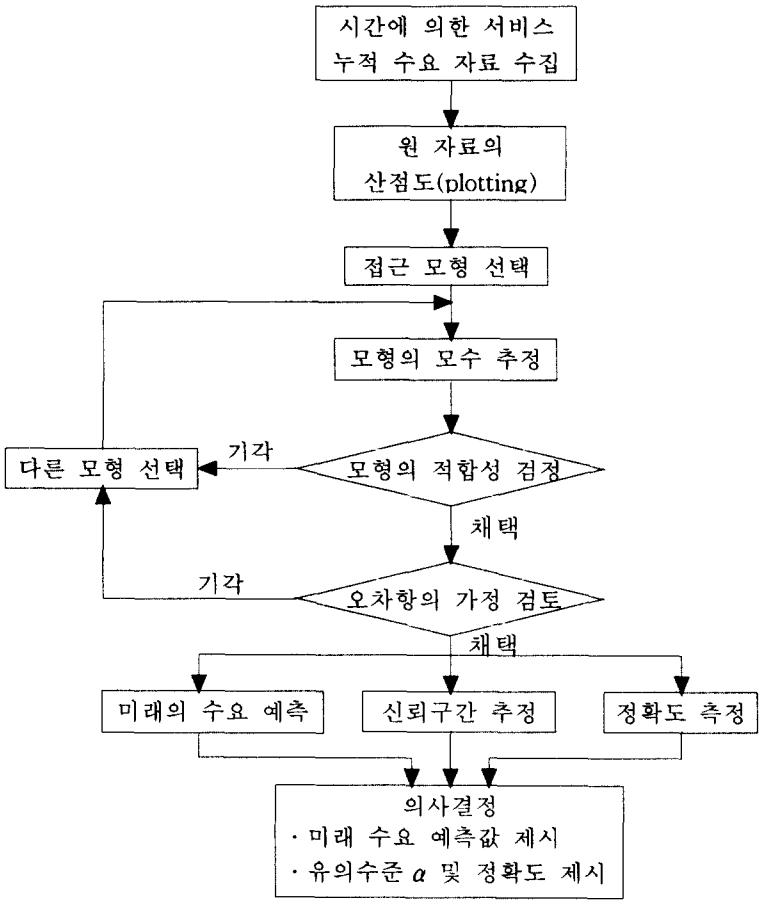
In order to determine procedures for appropriate model selection of technological growth curves, numerous time series that were representative of growth behavior were collected according to data characteristics. Three different growth curve models were fitted onto data sets in an attempt to determine which growth curve models achieved the best forecasts for types of growth data. The analysis of the results gives rise to an approach for selecting appropriate growth curve models for a given set of data, prior to fitting the models, based on the characteristics of the goodness of fit test.

1. 서론

시계열 자료에 대한 분석은 흔히 Box-Jenkins의 ARIMA 모형을 적용하여 분석하게 된다. 하지만, 시간에 따라 얻어진 자료가 누적 수요 자료일 때 적용하는 성장곡선 모형은 크게 선형모형과 비선형모형으로 분류된다(Young, 1993). 선형모형은 대부분 최대 극한값 K를 이미 아는 경우에만 분석이 가능하다. 그러나 최대 극한값을 모르고서 행하는 선형모형에 의한 분석은 많은 오차를 포함한다. 그것은 선형모형을 이용하여 분석할 때 최대 극한값을 미리 상정하거나 또는 가정을 하고서 분석하기 때문이다. 결국, 최대 극한값 K에 대한 신뢰성이 떨어지므로 수요 예측시 많은 오차를 포함하게 된다.

본 논문에서는 비선형 성장곡선 모형에 대해서만 다룬다. 비선형 성장곡선 모형으로는 로지스틱 모형(Logistic model), 프로빗 모형(Probit model), 곰페르츠 모형(Gompertz model) 등이 널리 사용된다(Young & Ord, 1990). 본 연구에서는 미국의 연도별 CATV 누적 가입자 수 자료를 이용하여 이 세 가지 모형에 각각 적합시켜 비교한다. 또, 모형선호의 기준으로 이용되는 제 통계량들을 이용하여 각 통계량 별로 위의 세 가지 모형을 비교한다.

비선형 모형선택의 기준으로 이용되는 통계량은 정확한 결정계수 R_c^2(Corrected Actual R-square), 정확한 수정결정계수 $\overline{R_c}^2$(Corrected Actual Adjusted R-square)를 이용하여 비교한다. 위의 결정계수는 흔히 선형회귀모형에서 얻어지는 결정계수와는 다르다. 이외의 모형선호기준 통계량은 <부록 2>를 참조한다(Wheelwright & Makridakis, 1985). 다음 <그림 1>은 성장곡선 모형에 의한 분석절차와 그 결과에 따른 의사결정을 나타내는 흐름도 이다.



< 그림 1 > 성장곡선 모형 분석 절차와 의사결정 흐름도

2. 비선형 성장곡선 모형

시간에 따라 얻어진 누적 수요 자료를 분석하기 위한 대표적인 비선형 성장곡선 모형은 아래와 같다(Young & Ord, 1990).

(1) 로지스틱 모형(Logistic Model)

$$Y_t = \frac{K}{1+\exp(\alpha+\beta t)} + \varepsilon_t \tag{2.1}$$

위 모형은 펄 성장곡선 모형(Pearl Growth Curve Model)으로 불리기도 한다.

(2) 프로빗 모형(Probit Model)

$$Y_t = K + \text{probnorm}\ (\alpha+\beta t) + \varepsilon_t \tag{2.2}$$

(3) 곰페르츠 모형(Gompertz Model)

$$Y_t = K\exp\{-\alpha\exp(-\beta t)\} + \varepsilon_t \tag{2.3}$$

위 비선형 성장곡선 모형에서 Y_t는 시간에 따른 누적 수요,

K는 최대 극한값(upper limit),
α, β는 추정할 모수(parameters),
$\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$을 가정

그러나 비선형 성장곡선 모형의 오차항(ε_t)들은 때때로 동일한 분포(identically distributed)를 따르지 않는 경우가 있다. 그 이유는 오차항의 분산이 등분산성을 충족하지 못하기 때문이다. 따라서 이러한 경우는 OLS(Ordinary Least Squares)보다 GMM(Generalized Method of Moments)에 의해서 모수를 추정하는 것이 보다 효율적인 것으로 알려졌다(Hansen, 1985).

이외의 극한값(K)을 이미 상정하거나 또는 가정하고서 이용하는 선형모형에 대한 것은 <부록 1>을 참조한다.

3. 모형선호기준 통계량

3.1 모형선호기준 통계량

비선형 성장곡선 모형에 대한 모형선호 기준으로 이용되는 통계량은 아래와 같다.

(1) 정확한 실제 결정계수(Corrected Actual R-square)

$$R_c^2 = 1 - \frac{SSE}{CSSA} \tag{3.1}$$

$$SSE = \sum_{t=1}^{N} (\hat{y}_t - y_t)^2$$

$$SSA = \sum_{t=1}^{N} y_t^2$$

$$CSSA = \sum_{t=1}^{N} y_t^2 - (\sum_{t=1}^{N} y_t)^2 = \sum_{t=1}^{N} (y_t - \bar{y})^2$$

y_t = 실제값(actual value),

$\hat{y}_t$ = 예측값(predicted value)

SSA는 실제(actual) 자료의 제곱합(Sum of the Squares)을 의미하고, CSSA에 있어서 $\bar{y}$는 실제 자료의 평균(actual mean)에 의해서 구해진다. 그리고 식 (3.1)은 흔히 선형회귀모형에서 구하는 결정계수와는 다르다.

(2) 정확한 실제 수정결정계수(Corrected Actual Adjusted R-square)

$$\overline{R_c}^2 = \frac{(n-1)\ R_c^2 - k}{n-k-1} \tag{3.2}$$

$k = p-1$,

n = 표본 수,

p = 추정된 모수의 수

결정계수는 모형의 설명력을 나타내는 것으로 결정계수의 범위는 0과 1사이의 값을 가지며 1에 근접한 값을 가질수록 좋은 모형이다.

(3) 예측의 정확성 통계량(a statistic measuring the accuracy of a forecast) $U1$

$$U1 = \frac{MSE}{\sqrt{\frac{1}{N}\sum_{t=1}^{N} y_t^2}} \tag{3.3}$$

(4) Theil's 부등식 계수(inequality coefficient) U

$$U=\frac{MSE}{\sqrt{\frac{1}{N}\sum_{t=1}^{N} y_t^2}+\sqrt{\frac{1}{N}\sum_{t=1}^{N} \hat{y}_t^2}} \tag{3.4}$$

식(3.3)과 (3.4)는 모형적용에 의해 얻어진 예측값의 정확성을 나타내는 통계량으로 U1과 U의 계수 값은 0에 근접할 때 모형 적용에 의한 예측이 정확한 것으로 판정한다. 특히, Theil's의 부등식 계수 값이 U =0.55 이하이면 예측이 정확한 것으로 판정한다(Lindberg, 1982; McNess, 1979).

3.2 오차항의 가정 검토

(1) 오차항의 독립성 검정

오차항의 독립성 검정은 더빈-왓슨(Durbin-Watson) 검정 통계량에 의해서 검토한다.

$$D=\frac{\sum_{i=1}^{n}(e_i-e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{3.5}$$

$$0<D<4$$

D 값은 0부터 4까지 가질 수 있으며 2에 가까울수록 오차 항들은 서로 독립이라고 보며, 0에 가까우면 양의자기상관이 있고, 4에 가까우면 음의자기상관이 있다고 본다. 이 경우 선형회귀분석이 아닌 시계열분석으로 한다. 그리고 누적수요 자료인 경우는 비선형 성장곡선 모형으로 분석해야 한다.

(2) 오차항의 등분산성 검정

오차항의 등분산성 검정은 오차항에 의한 잔차 플롯으로 검토한다. 그리고 이분산성 일 때는 OLS(Ordinary Least Squares)에 의한 방법보다는 GMM(Generalized Method of Moments)에 의해서 모수를 추정한다.

(3) 오차항의 정규성 검정

오차항의 정규성 검정은 샤피로-윌크(Shapiro-Wilk) 검정 통계량과 P-P 플롯으로 검토한다.

4. 사례 연구

4.1 본 연구에 이용된 자료

아래 표의 자료는 년간 미국 CATV 누적 가입자 수(한국통신학회지 1994년 제11권 11호 p.59)를 나타낸다.

< 표 1 > 미국 CATV 년간 누적 가입자수

년도	가입자수	년도	가입자수	년도	가입자수
1970	4,498,030	1978	13,391,910	1986	42,237,140
1971	5,569,810	1979	14,814,380	1987	44,970,880
1972	6,484,380	1980	17,671,490	1988	48,636,520
1973	7,163,340	1981	23,219,200	1989	50,897,080
1974	8,230,310	1982	29,340,570	1990	53,900,000
1975	9,196,690	1983	34,113,790	1991	56,000,000
1976	10,787,970	1984	37,290,870	1992	57,868,170
1977	12,168,450	1985	39,872,520	1993	59,397,390

4.2 비선형 성장곡선 모형의 비교

< 표 2 > 모수 추정 통계량

통계량 / 모형	K (최대 극한값) Prob> | T |	α Prob> | T |	β Prob> | T |
로지스틱 모형	66375427 P-value=0.0001 **	3.013628 P-value=0.0001 **	0.224129 P-value=0.0001 **
프로빗 모형	68990136 P-value=0.0001 **	-1.784128 P-value=0.0001 **	0.128710 P-value=0.0001 **
곰페르츠 모형	90140521 P-value=0.0001 **	3.831498 P-value=0.0001 **	0.099617 P-value=0.0001 **

<표 2>에서 세 가지 모형에 대한 모수 추정 결과 P값이 작으므로 통계적 유의성을 확보한다.

< 표 3 > 모형선호기준 통계량[1)]

통계량 / 모형	ME	MPE	MAE	MAPE	MSE	RMSE	RMSPE
로지스틱 모형	-13779015	-37.7827	13785980	37.93750	3.17323E14	17813550	41.7955
프로빗 모형	28325210	316.3553	31635044	322.05178	1.45591E15	38156361	503.9534
곰페르츠 모형	17606142	112.7719	17606142	112.7719	3.33416E14	18259691	137.9210

<표 3>의 통계량은 모형선호 기준 통계량으로 비교적 작은 값을 가지는 것이 좋은 모형으로 선택된다. 위 <표 3>에서 로지스틱 증가곡선 모형이 가장 작은 값으로 나타나므로 좋은 모형으로 선호된다.

< 표 4 > 모형선호기준 통계량

통계량 / 모형	R_c^2	$\overline{R_c}^2$	$U1$	U
로지스틱 모형	0.9942	0.9936	0.5170	0.3468
프로빗 모형	0.9925	0.9917	1.1021	0.4145
곰페르츠 모형	0.9892	0.9881	0.5299	0.2170

<표 4>의 정확한 실제 결정계수 범위는 0과 1 사이의 값을 가지며 1에 가까운 값을 가질 때 좋은 모형으로 선택된다. <표 4>에서 로지스틱 성장곡선 모형이 R_c^2과 $\overline{R_c}^2$의 결정계수 값이 가장 높다. 그리고 모형 예측의 정확성을 나타내는 통계량 $U1$과 Theil's 부등식 계수 U에 있어서도 기준값 0.55보다 작은 값을 갖는다. 위 결과에 의해 로지스틱 성장곡선 모형이 가장 좋은 모형으로 선택됨을 알 수 있다. 또한 예측 정확성 통계량 값이 $U1$=0.517로써 정확히 예측된 것으로 판정된다.

4.3 오차항의 가정 검토

4.3.1 오차항의 독립성 검정

< 표 5 > 오차항의 독립성 검정 통계량

	로지스틱 모형	프로빗 모형	곰페르츠 모형
DW 통계량 값	0.421	0.358	0.290

1) 통계량 값들은 부록 2와 앞 3.1절의 식에 의하여 SAS를 이용하여 구한 것임.

<표 5>에서 더빈-왓슨 검정통계량 값이 0에 가까운 값을 가지므로 선형회귀모형에 의한 최소자승법(OLS)에 의해서 모수를 추정시 독립성이 위배됨을 예견할 수 있다.

4.3.2 오차항의 등분산성 검정

세 가지 비선형 증가곡선 모형에 대한 오차항의 가정에 대한 잔차 플롯 결과 모두 등분산성 가정을 충족한 것으로 나타났다.

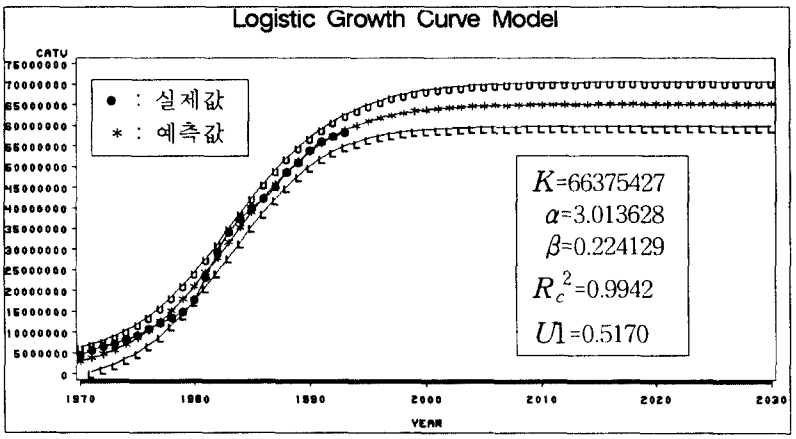
4.3.3 오차항의 정규성 검정

< 표 6 > 오차항의 정규성 검정 통계량

	로지스틱 모형	프로빗 모형	곰페르츠 모형
샤피로-윌크 통계량 값	$\hat{W}=0.94325$ $P_r< W=0.1982$	$\hat{W}=0.93198$ $P_r< W=0.1105$	$\hat{W}=0.84275$ $P_r< W=0.0015$

오차항의 정규성 가정에 대한 검정 결과 곰페르츠 모형만이 위배됨을 알 수 있다.

4.4 비선형 성장곡선 모형에 의한 예측과 신뢰구간 추정



< 그림 2 > 로지스틱 성장곡선 모형에 의한 예측과 신뢰구간 추정

<그림 2>는 로지스틱 성장곡선 모형에 의한 예측과 95% 신뢰구간 추정선을 나타낸 것이다. 앞서 오차항의 가정 검토와 모형선호기준 통계량을 비교한 결과 로지스틱 성장곡선 모형이 자료에 대한 예측 설명력이 가장 높았다. 그리고 현재 미국의 경우 CATV 가입자는 전체 가입자 수(최대 변곡점: K=66375427)에 대해서 비교하여 보면

앞으로는 증가 추세가 매우 적음을 알 수 있다. 그것은 가입자 예측선이 최대 변곡점을 향해 거의 수평을 유지하기 때문이다.

5. 결론 및 토의

세 가지 비선형 성장곡선 모형에 대하여 분석한 결과를 살펴보면, 처음 각 개별 검정에서는 세 가지 모형에 대한 모든 결과가 통계적 유의성을 가진다. 그리고 모형선호기준을 위한 각 통계량 값들을 비교하여 보면 로지스틱 모형이 가장 좋은 모형으로 판정된다. 그것은 모형선호기준 통계량 값들이 다른 모형에 비하여 비교적 작은 값을 가지며, 결정계수에 의한 통계량 값은 가장 큰 값을 갖는다. 또한 예측의 정확성을 나타내는 $U1$과 U 통계량 값이 기준값 0.55 보다 작은 값을 가지므로 정확히 예측되었다. 그리고 세 가지 모형을 이용한 자료분석은 SAS/STAT NLIN 명령어와 SAS/ETS MODEL 명령어를 이용한다. 이때, NLIN 명령어를 이용하면 모형에서 얻어진 예측값과 신뢰대를 OUTPUT에 저장하여 바로 SAS/GRAPH를 이용한 그래프 분석이 가능하다. 그리고 MODEL 명령어에서는 SOLVE 옵션으로 모형선호기준 통계량들을 쉽게 구하는 장점이 있다.

비선형 성장곡선 모형에 대한 분석은 쉽지가 않다. 그것은 비선형 모형에 대한 최대 극한값과 그 외 모수 추정을 하는데 있어서 수치해석 방법에 의한 초기 값 설정에 따라 모수가 쉽게 추정되지 않기 때문이다. 이러한 어려움 때문에 선형 변환을 하여 최대 극한값을 임의로 상정하거나 또는 가정 하에 구하는 경우가 흔하다. 이 경우 오차항의 가정 검토 없이 추정된 결과치를 그대로 받아들인다면 많은 오류를 포함하게 된다. 또한, 최대 극한값이 과학적 증거 사실이 없기 때문이다. 따라서 본 논문은 실제 응용자료를 가지고서 최적의 비선형 성장곡선 모형을 찾는 과정과 모형선택을 위한 여러 개의 통계량 값과 흔히 선형회귀모형에서 이용되는 결정계수와는 다른 결정계수 R_c^2과 $\overline{R_c}^2$을 이용하여 비교한 후 최적 모형을 선택하였다. 그리고 $U1$과 U 통계량 값을 이용하여 예측의 정확성을 비교하였다.

참고문헌

[1] Bewley, R. and Fiebig, D.G.(1988), "A Flexible Logistic Growth Model with Application in Telecommunications," *International Journal of Forecasting,* Vol. 4, pp. 177-192.

[2] Hansen, L.P.(1985), "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators," *Journal of Econometrics*, Vol. 30, pp. 203-238.

[3] Lancaster, G.A. and Wright, G.(1983), "Forecasting the Future of Video Using a Diffusion Model," *European Journal of Marketing,* Vol. 17(2), pp. 70-79.

[4] Lindberg, B.c.(1982), "International comparison of growth in demand for a new durable consumer product," *Journal of Marketing Research,* Vol. 19, pp. 364-371.

[5] Makridakis, S. and Hibon, M.(1979), "Accuracy of forecasting: an empirical investigation(with discussion)," *The Journal of the Royal Statistical Society, Series A,* Vol. 142, pp. 97-145.

[6] Makridakis, S.(1988), "Metaforecasting : Ways of Improving Forecasting Accuracy and Usefulness," *International Journal of Forecasting,* Vol. 4, PP. 467-491.

[7] McNess, S.K.(1979), "The forecasting record for the 1970's," *New England Economic Review,* September-October.

[8] Meade, N.(1985), "Forecasting using Growth Curves : An Adaptive Approach," *Journal of the Operational Research Society,* Vol. 36, pp. 1103-1115.

[9] SAS/Procedure and ETS Guide for PC, Ver. 6Ed.(1990), SAS Institute Inc.

[10] SAS/Procedure and STAT Guide for PC, Ver. 6Ed.(1990), SAS Institute Inc.

[11] Sharif, M.N. and Kabir, C., "A Generalized Model for Forecasting Technological Substitution," *Technological Forecasting and Social Change,* Vol. 8, pp. 353-364.

[12] Wheelwright, S.C. and Makridakis, S.(1985), *Forecasting Methods for Management,* Fourth Edition, John Wiley & Sons, New York.

[13] Young, P. and Ord, J.K.(1990), "Model Selection and Estimation for Technological Growth Curves," *International Journal of Forecasting,* Vol. 5, pp. 501-514.

[14] Young, P.(1993), "Technological Growth Curves: A Competition of Forecasting Models," *Technological Forecasting and Social Change,* Vol. 44, pp. 375-389.

부록 1. 선형 모형

(1) 맨스필드-블랙맨 모형(Mansfield-Blackman Model)

$$\ln\left(\frac{Y_t}{K-Y_t}\right)=\beta_0+\beta_1 t+\varepsilon_t \tag{1.1}$$

위 모형은 피셔-프라이 모형(Fisher-Pry Model)으로 불리기도 한다.

(2) 선형 곰페르츠 모형(Linear Gompertz Model)

$$\ln\left(-\ln\left[\frac{Y_t}{K-Y_t}\right]\right)=\beta_0+\beta_1\ln t+\varepsilon_t \tag{1.2}$$

(3) 와이블 모형(Weibull Model)

$$\ln\left(\ln\left[\frac{K}{K-Y_t}\right]\right)=\beta_0+\beta_1\ln t+\varepsilon_t \tag{1.3}$$

(4) 베스 모형(Bass Model)

$$y_t=\beta_0+\beta_1 Y_{t-1}+\beta_2(Y_{t-1})^2+\varepsilon_t \tag{1.4}$$

(5) NSRL(Nonsymmetric Responding Logistic Model)

$$\ln y_t=\beta_0+\beta_1\ln(Y_{t-1})+\beta_2\ln(K-Y_{t-1})+\varepsilon_t \tag{1.5}$$

(6) 하베이 모형(Harvey Model)

$$\ln y_t=\beta_0+\beta_1 t+\beta_2\ln(Y_{t-1})+\varepsilon_t \tag{1.6}$$

K는 최대 극한값(upper limit),
$\beta_0, \beta_1, \beta_2$는 추정할 모수(parameters),
$\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$을 가정

위 선형 모형에서 베스 모형과 하베이 모형을 제외한 나머지 모형들은 최대 극한값 K 값을 아는 경우에만 선형 모형에 의한 추정이 가능하다.

부록 2. 모형선호기준 통계량

(1) 평균오차(Mean Error)

$$ME=\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t-y_t) \tag{2.1}$$

(2)평균백분율오차(Mean Percentage Error)

$$MPE=\frac{100}{N}\sum_{t=1}^{N}(\hat{y}_t-y_t)/y_t \tag{2.2}$$

(3) 평균절대오차(Mean Absolute Error)

$$MAE=\frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t-y_t| \tag{2.3}$$

(4) 평균절대백분율오차(Mean Absolute Percentage Error)

$$MAPE=\frac{100}{N}\sum_{t=1}^{N}|(\hat{y}_t-y_t)/y_t| \tag{2.4}$$

(5) 평균제곱오차(Mean Squared Error)

$$MSE=\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t-y_t)^2 \tag{2.5}$$

(6) 평방근평균제곱오차(Root Mean Square Error)

$$R\,MSE=\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t-y_t)^2} \tag{2.6}$$

(7) 평방근평균제곱백분율오차(Root Mean Square Percentage Error)

$$R\,MSPE=100\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t-y_t/y_t)^2} \tag{2.7}$$