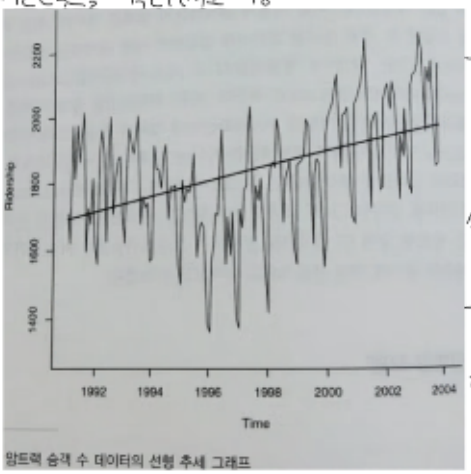
17장 회귀분석을 기반으로 한 예측

1. 추세를 반영한 모델

- 선형추세

선형추세를 반영할 수 있는 선형회귀모델을 만들기 위해 관측값(혹은 관측값들의 함수)을 반응변수(Y)로

시간인덱스를 예측변수(X)로 사용



암트랙 승객 수 데이터의 선형 추세 그래프

→ 전반적인 추세는 선형을 따르고 있지 않음.

그러나 이 데이터를 통해 선형추세모델이 어떻게 구축되는지

살펴보고, 향후 더 적합한 모델에 대해 논의

→ 승객 수와 시간 사이의 관계를 설명하는 선형식을 구축하기위해

승객수를 반응변수(Y)로, 시간인덱스 ($t=1,2,3,\cdots$)를 예측변수 X로 설정

└ 예측변수가 시간인덱스 하나이므로 다음과 같은 단순회귀모델로 표현가능 $Y_t = \beta_0 + \beta_1 t + \varepsilon$

Y_t - t 시점에서의 승객 수 ε - 선형 회귀모델에서 가정하는 표준오차 → 시계열 구성요소 중 수준(β_0), 추세(β_1), 잡음(ε)을 반영

계절변동은 포함하고 있지않음

회귀모델결과 해석 시 주의사항 : 단순히 추정된 회귀식의 계수와 통계적 유의성만을 모델 평가 기준으로 삼는 것은 잘못된 결론을 도출할 수 있음.

└위 그래프에서 전체적인 추세가 선형이 아님에도 불구하고 회귀식의 계수를 통계적으로 검증하면 선형추세가 적합한 것으로 판정됨

└ 학습데이터의 표준오차는 특정 추세에 모델이 얼마나 적합한지에 대한 측도는 될 수 있지만 모델의 예측 성능을 평가하는데는 모순이 있을 수 있음

→ 이 경우 검증데이터의 실제값과 모델로부터 얻은 예측값의 차이를 살펴봤 것이 바람직한 방법

- 지수추세

선형회귀모델을 사용하면 선형 추세뿐 아니라 몇몇 다른 형태의 추세들 또한 모델링 가능

└ 대표적인 것이 지수추세 - 지수추세는 시간이 흐름에 따라 곱의 형태로 증가 혹은 감소 ($Y_t = ce^{\beta_1 t + \varepsilon}$) 하는 시계열 패턴

지수추세를 반영한 회귀모델을 수립하기 위해서는 반응변수 Y를 $\log(Y)$로 대체하고 선형회귀모델을 구축해야함 $\log(Y) = \beta_0 + \beta_1 t + \varepsilon$

└ 지수추세를 반영하는 회귀모형은 주로 기업의 판매실적의 성장추이를 분석하는데 널리 활용됨

- 일반적으로 선형회귀분석에서 반응변수가 다른 경우 모델의 예측 정확도를 비교하기 위해서는 반드시 단위일치 필요.

└ 선형추세모델은 Y를 기준으로 수립, 지수회귀모델은 $\log(Y)$를 기준으로 수립

⇒ 두 모델의 예측성능을 비교하기 위해서는 단위통일 필요.

- 다항추세

선형회귀분석을 통해 구현할 수 있는 또 다른 비선형 추세 - 다항추세

다항추세의 특별한 경우인 이차추세식 $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon$ 은 예측변수 t^2을 추가함으로써 얻을 수 있음.

일반적으로 어떤 형태의 추세라도 수학적 표현이 가능 - 그러나 현 데이터의 적합 성능만을 고려하여 너무 복잡한 형태의 모델 고집 X
⇒ 과적합의 위험 초래 (이러한 문제를 피하기 위해 검증데이터를 통한 성능 검증이 필수)

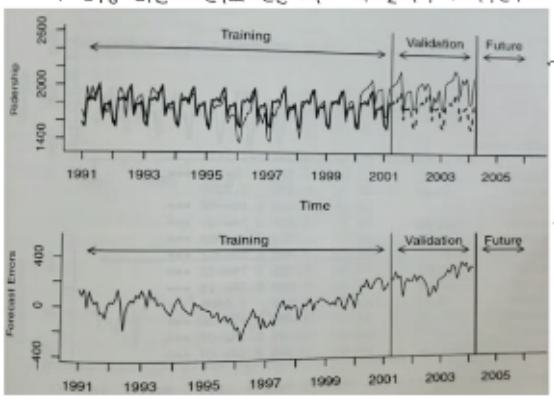
2. 계절변동 반영모델

시계열에서 계절패턴이 존재 - 특정 계절의 관측치가 다른 계절에 비해 일정하게 크거나 작은 값을 가지는 것을 의미.

└ 시계열에서 주별로 패턴이 반복되거나 혹은, 분기별로 반복되는 패턴을 보일 때 계절변동이 존재한다고 말함

회귀모델에 계절변동을 반영하기 위해서는 계절을 표현할 수 있는 서로소 범주형 변수 생성 필요.
→ '월'을 보고싶다면 11개의 가변수생성

└ 이 범주형 변수는 가변수로 변환되어 회귀모델 구축 시 예측변수로 사용됨.



→ 예측 값과 실제 값 그래프와 단차

└ 모델이 데이터의 계절변동을 잘 반영하고 있는 것처럼 보이지만 추세를 반영한 모델에 포함되어 있지 않기 때문에 전반적으로 만족스러운 결과가 아님을 볼 수 있다.

위에서 언급한 이진변수를 사용하여 구축된 회귀모델은 계절변동의 합(Additive)을 반영 ← 반응변수 값의 평균이 기준 월과 비교해볼 때, 특정수치만큼 크거나 작다는 것을 의미

회귀모델을 통해 특정 월이 기준 월에 비해 몇 % 증가(혹은 감소) 했는지와 같은 비율을 알아보기 위해서는 계절변동의 곱(multiplicative)을 반영하는 모델 이용해야함.

3. 추세와 계절변동을 반영한 모델

추세+계절변동을 동시에 반영하는 모델 ⇒ 추세 모델과 계절변동 모델을 통합함으로써 얻을 수 있음.

4. 자기상관과 아리마 모델

전통적인 회귀모델 - 관측치 간 서로 독립을 가정

└ 그러나 시계열 관측치들은 대부분 서로 상관관계를 갖고 있다. → 이러한 관측치 간의 상관관계를 자기상관관계라고 함.

↳ 이 자기상관정보를 활용하면 보다 더 정확한 예측을 할 수 있다. → 상관관계에 따라 예측값 조정 가능

- 자기상관의 계산

자기상관 - 시계열 데이터에서 인접한 관측치 간의 관계

↳ 두 개의 변수 사이의 관계를 나타내는 일반적인 상관관계와는 달리 하나의 시계열 변수와 그 자신의 관계정도만 보여줌

자기상관은 시계열과 시차를 둔 그 자신 시계열 사이의 상관계수를 구함으로서 얻을 수 있음.

↳ 1 혹은 그 이상의 기간만큼 차이를 두고 원래 시계열을 그대로 옮긴 것.

몇몇 전형적인 자기상관관계의 특징

- 지연이 1보다 클 때의 강한 자기상관(양/음) – 데이터의 주기적인 패턴을 의미 (시차)
- 1-지연 양의 자기상관 – 연속적인 값들이 전반적으로 같은 방향성을 지니고 있음을 의미. 강한 선형추세가 존재할 때, 1-지연 양의 자기상관이 강하게 나타난다.
- 1-지연 음의 자기상관 – 시계열의 변동이 심할 때 나타남. 어떤 기간이 큰 값을 가지면 다음 기간은 작은 값을 가지고, 반대로 작은 값을 가지면 다음 기간에 큰 값이 나타나는 경우가 이에 해당.

시계열의 자기상관관계를 통해 계절변동의 패턴을 찾아낼 수 있다. (ex. 6-지연 마다 음의 상관 ⇒ 분기마다 바뀌는 패턴)

본 시계열의 자기상관과 더불어 잔차 시계열의 자기상관 정보도 유용하게 사용될 수 있다.

↳ 예측모델이 수립되면, 예측 값과 실제 값의 차이인 잔차를 구할수가 있고 이를 이용하여 잔차의 자기상관을 계산할 수 있다.

만약, 구축된 모델이 계절적 변동을 잘 반영하고 있다면 잔차를 이용한 자기상관 그래프에는 더 이상 계절적인 패턴이 보이지 않을 것임.

- 자기상관 정보가 포함된 예측의 개선

일반적으로 자기상관 정보를 활용하는 방법 ┌ 회귀모델에 자기상관 정보를 바로 포함시키는 방법
                                              └ 잔차값을 이용하여 2차 예측모델을 만드는 것

자기상관 정보가 직접 포함되어 있는 모델의 대표적인 것 – 아리마 (ARIMA ; Autoregressive Integrated Moving Average)

ARIMA 모델의 특수한 형태인 자기회귀 (AR, Autoregressive) 모델은 일반적인 선형회귀모델과 비슷.

ex) 2차 자기회귀모델 (AR(2))

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon$$

↳ 반응변수의 과거 기간 값들이 예측변수로 사용된다는 점에서 차이.

모델을 구축하는 과정은 1-, 2-지연 시계열을 예측변수로 포함하여 선형 회귀분석을 수행하는 것과 유사.

↳ 그러나 모델의 계수를 추정하는 방식이 선형회귀모델에서 사용하는 최소제곱법이 아니라 아리마 모델에서 사용하는 추정법을 사용

아리마 모델은 시계열 데이터의 평균과 분산이 시간에 관계없이 일정한 정상 시계열(stationary)을 가정하기 때문에

추세나 계절변동이 존재하는 비정상 시계열(nonstationary)의 경우에는 정상 시계열로의 변환작업이 필요.

↳ 아리마 모델은 강건하지않고, 많은 경험과 통계적 지식을 필요로 하기 때문에 일반적으로는 널리 사용되지않음.

쉽게 예측에 활용할 수 있는 특별한 AR 모델 – 이는 단기예측에 있어 효과적

① 예측모델을 통해 k-기간 앞선 예측값(F_{t+k})을 구한다.

② AR(혹은 다른) 모델을 통해 k-기간 앞선 예측값의 오차(E_{t+k})를 구한다.

③ 앞서 구한 k-기간 앞선 예측값에 오차를 보정하여 새로운 예측값을 구한다. $F'_{t+k} = F_{t+k} + E_{t+k}$

특히 낮은 차수의 AR모델을 잔차시계열(혹은 예측오차)에 적용함으로서 미래의 예측오차를 정확히 예측할 수 있다.

또한 잔차시계열은 자기상관관계 외 다른 추세나 주기적 변화를 갖지 않을것으로 간주되기 때문에 데이터 변환 작업이 따로 필요.

AR 모델을 잔차시계열에 적합할 때는 먼저 잔차의 자기상관을 확인한다. 자기상관관계의 확인을 통해 나타난 지연정도에 따라 자기회귀모형의 차수 결정

잔차시계열의 자기상관관계가 1-지연일 경우 자기상관성도가 크게 나타났다면 AR(1) 모델이 적합: $E_t = \beta_0 + \beta E_{t-1} + \varepsilon$

↳ 시간 t에서의 잔차를 의미

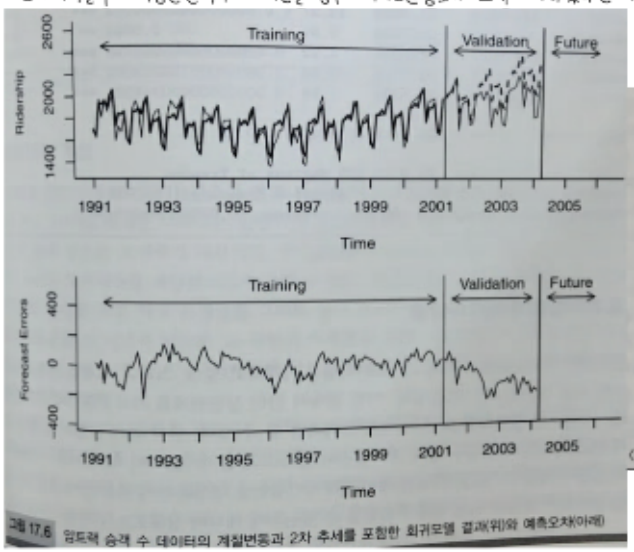
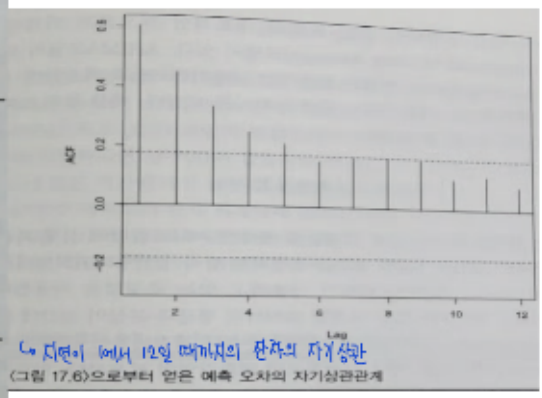


그림 17.6 임의적 승객 수 데이터의 계절변동과 2차 추세를 포함한 회귀모델 결과(위)와 예측오차(아래)



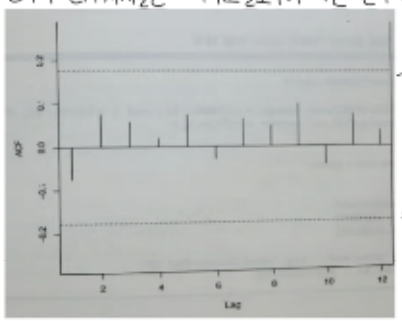
↳ 지연이 1에서 12일 때까지의 잔차의 자기상관
(그림 17.6)으로부터 얻은 예측 오차의 자기상관관계

↳ 1-지연일 때 자기상관이 가장 크게 나타났으므로 AR(1) 모델이 적합

AR(1)모델이 적합한 이유는 한 기간 값들 사이에 상관이 있으면, 이 상관관계는 2-기간, 3-기간, 혹은 더 많은 기간에 까지 영향력을 미치기 때문임.

끝으로 주어진 시계열의 자기상관관계로부터 필요한 정보를 모두 활용했는지, 혹 시계열에 더 이상의 유의한 정보가 남아있는지 여부를 확인하기위해서 잔차로부터 파생된 또 다른 시계열의 자기상관을 확인할 수 있다.

잔차의 잔차시계열은 회귀모델로부터 나온 잔차에 AR(1)모델을 적용한 후에 생기는 잔차의 시계열.



→ 잔차의 잔차에 대한 자기상관

더 이상의 자기상관이 존재하고 있지 않음

⇒ AR(1) 모델이 자기상관의 정보를 적절하게 반영했다는 것을 의미

AR 모델을 추가함으로써 예측의 정확도를 개선하고 이는 단기예측에 적합.

↳ 차수가 k인 자기회귀모델은 향후 k기간의 예측에만 유용하기 때문임. k기간보다 더 먼 예측은 실제 값보다 앞서 예측한 값에 더 많은 영향을 받게 됨. ex) 2001년 3월에 5월의 잔차를 예측하려면 4월의 잔차가 필요. 그러나 4월의 값은 없기 때문에 예측값으로 대체. ⇒ 5월의 예측값은 4월의 예측값을 기반으로 구해짐.

- 예측성 검증

예측성능을 평가할 수 있는 유용한 방법 중 하나는 시계열이 확률보행과정을 따르고 있는 지 여부를 확인해보는 것.

↳ 확률보행이란 시계열이 특정 시간에서 다음 시간까지 무작위로 변화하는 현상

확률보행은 AR(1) 모델의 특별한 경우 : 기울기 계수가 1

$Y_t = \beta_0 + Y_{t-1} + \varepsilon_t \Rightarrow Y_t - Y_{t-1} = \beta_0 + \varepsilon_t$

→ 시간 t-1과 t 사이의 값 차이는 원의의 어떤 값도 될 수 있다. 기본적으로 위 식을 통해 얻은 예측값은 가장 최근 관측값이 되며 다른 어떤 정보도 반영되지 않음.

시계열 자료가 확률보행과정을 따르는 지 여부 ⇒ AR(1) 모델을 구축하고, 그 계수의 유의성을 확인하는 가설검정을 수행 ($H_0: \beta_1 = 1$ vs $H_1: \beta_1 \neq 1$)

⇒ 확률보행을 따른다면 앞서 말한 어떤 형태의 예측방법도 적용되기 어려움.