**4과목 4절**

4절 회귀분석 (★★★개념 ★★★ R Output 해석 3-5문제!!!)

1. 회귀분석 개요

정의 : 하나나 그 이상의 독립 변수들이 종속변수에 미치는 영향을 추정 할 수 있는 통계기법   
독립변수의 개수가 하나이면 단순 선형회귀분석, 독립변수의 개수가 두개 이상이면 다중선형회귀분석

회귀분석 변수 :   
- 영향을 받는 변수 y : 반응변수, 종속변수, 결과변수  
- 영향을 주는 변수 x : 설명변수, 독립변수, 예측 변수   
선형 회귀 분석의 가정  
1) 선형성 : 입력변수와 출력변수의 관계가 선형이다  
2) 등분산성 : 오차의 분산이 입력변수와 무관하게 일정하다. 잔차플롯 (산점도)를 활용하여 잔차와 입력변수 간에 아무런 관련성이 없게 무작위적으로 고루 분포 되어야 등분산성 가정을 만족하게 된다

3) 독립성 : 입력변수와 오차는 관련이 없다. 자기상관 (독립성)을 알아보기 위해 Durbin-Waston 통계량을 사용하며 주로 시계열 데이터에서 많이 사용  
5) 비상관성 : 오차들끼리 상관성이 없다  
6) 정상성 (정규성) : 오차의 분포가 정규분포를 따른다. Q-Qplot, Kolmogolove-Smirnove 검증, Shaprio-Wilk 검정을 활용하여 정규성을 확인

1. 단순선형회귀분석   
   하나의 독립변수가 종속변수에 미치는 영향을 추정 할 수 있는 통계기법

회귀분석에서 검토사항 !!!   
√ 회귀 계수들이 유의미한가 ? 회귀 계수의 t 통계량의 p value가 0.05보다 작으면 통계적으로 유의미

√ 모형이 얼마나 설명력을 갖는가 R2 값을 통해 확인. 높을수록 설명력이 높다  
√ 모형 데이터를 잘 적합하고 있는가. 잔차를 그래프로 그리고 회귀진단

1. 다중선형회귀분석 (다변량 회귀분석)

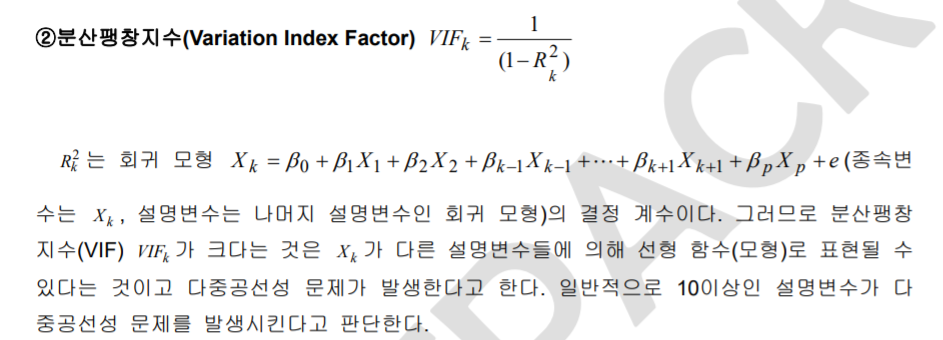
√ 모형의 통계적 유의성 : F 통계량으로 확인, p value가 0.05 보다 작으면 추정된 회귀식이 통계적으로 유의미하다고 볼 수 있다  
√ 회귀 계수의 유의성 : t 통계량 p value로 확인

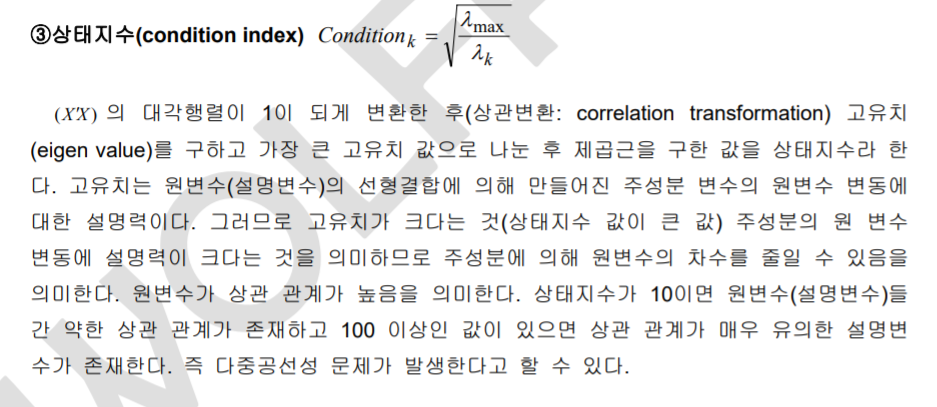
√ 모형의 설명력 : R^2, 수정된 결정 계수로 확인  
√ 모형의 적합성 : 모형이 데이터를 잘 적합하고 있는지 잔차, 종속변수의 산점도로 확인

√ 데이터가 전제하는 가정을 잘 만족 시키는가 : 선형성, 독립성, 등분산성, 비상관성, 정상성

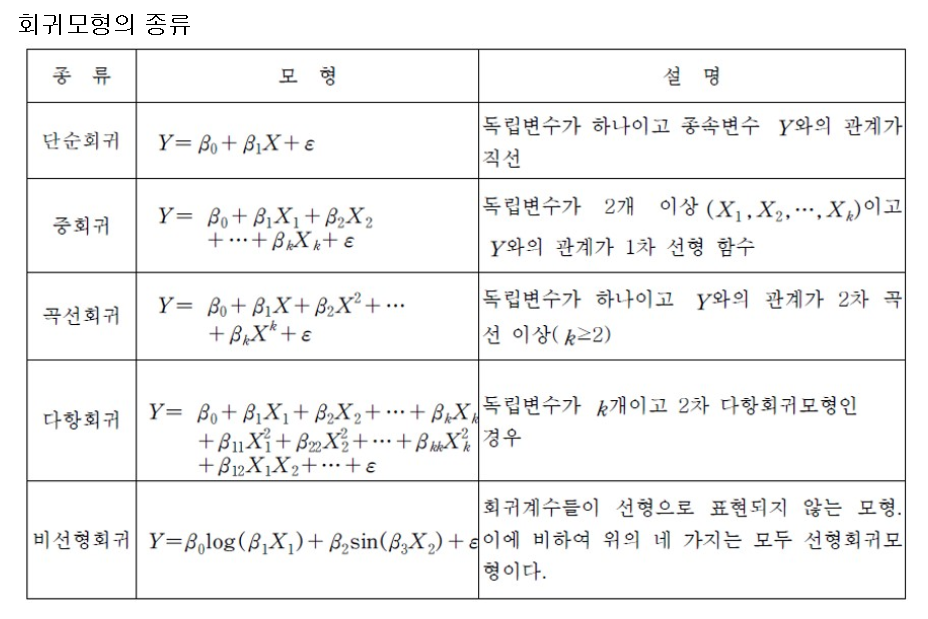
√ 다중공선선성 : 다중 회귀 분석에서 설명 변수들 사이에 선형관계가 존재하면 회귀 계수의 정확한 추정이 곤란하다. Check 방법  
 1) 분산 팽창 요인 (VIF) : 4보다 크면 다중공선성 존재. 10보다 크면 심각한 문젴

2) 상태지수 : 10 이상이면 문제가 있다고 보고, 30 이상이면 심각한 문제   
다중공선성이 존재하면, 문제가 있는 변수를 제거 하거나, 주성분 회귀, 능형회귀 무형을 적용하여 문제 해결





1. 회귀 분석의 종류   
   추가로 로지스틱 회귀 : 종속변수가 범주형인 경우에 적용



1. 회귀분석사례
2. 회귀분석사례
3. 최적회귀방정식  
   [최적회귀방정식의 선택]  
   3) 단계적 변수선택

* 전진선택법 : 절편만 있는 상수 모형으로부터 시작해 중요하다고 생각되는 설명변수부터 차례로 모형에 추가. 변수 개수가 많을 때도 사용가능 하지만, 변수의 작은 변동에도 결과가 크게 달라지는 안정성이 부족
* 후진제거법 : 후보 모두를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하고 더 이상 제거 할 변수가 없을 때 모형을 선택. 변수의 개수가 많을 때 사용하기 어려움
* 단계선택법 : 전진선택법에 의해 변수를 추가하면서 새롭게 추가된 변수에 기인해 기존 변수의 중요도가 약화 되면 해당 변수를 제거하는 등 단계별로 추가, 또는 제거

[벌전화된 선택기준]  
개요 : 모형의 복잡도에 벌점을 주는 방법으로, AIC 방법과 BIC 방법이 주로 사용  
방법 : AIC (Akaike information criterion), BIC (Bayesian information criterion)  
설명 : AIC, BIC를 계산하고 그 값이 최소가 되는 모형!  
모든 후보 모형들에 대해 AIC 또는 BIC 를 계산하고 그 값이 최소가 되는 모형을 선택.   
모형선택의 일치성 : consistency inselection, 자료의 수가 늘어날 때 참인 모형이 주어진 모형 선택 기준의 최소값을 갖게 되는 성질   
이론적으로 AIC 에 대해서 일치성이 성립하지 않지만, BIC 는 주요 분포에서 이러한 성질이 성립  
AIC를 활용하는 방법이 보편화된 방법  
그 밖의 벌점화 선택 기준으로 RIC, CIC, DIC 가 있다.