[4주차]일변량, 이변량

일변량의 자료탐색

일변량 데이터((univariate variable) :

각 단위에 대해 하나의 속성만 측정하여 얻어진 변수의 측정값(단변량 데이터)

수집된 자료의 대푯값이나 변동의 크기 등을 요약하여 특정한 수치로 나타내는 통계를 "기술통계량(descriptive statistics)" ⇒ 데이터(변수)의 특성 파악 가능

평균 : 데이터 중심위치의 측도

중앙값 : 전체 자료값을 가장 작은 값에서 큰 값으로 크기순서로 배열하여, 가운데 위치하는 값

분산 : 평균으로부터 흩어져 있는 측도

표준편차 : 데이터와 같은 단위를 갖는 산포의 측도

편차 : 데이터의 중심위치가 평균이라면, 관측값과 평균의 차이

범위 : 데이터의 최댓값 - 최소값

사분위수범위 : 데이터의 상하 25%의차이 ⇒ 3사분위수 - 1사분위수 = 사분위수 범위

이변량 데이터의 탐색

두 연속형 변수가 서로 짝을 이루었을 때(혹은 서로 의존적일 때)

두 변수는 서로 관계가 있는가 ? ⇒ 상관분석

관계가 있다면, 두 변수는 어떤 관계가 있다고 말할 수 있겠는가 ? ⇒ 회귀분석

산점도 평활

두 변수의 관계가 항상 직선의 경향만이 있는것이 아님

두 변수 사이의 관계에서 직선의 관계로 설명되지 못한 경우 산점도 평활(scatterplot smoothing)방법을 사용

평활은 window의 크기와 가중최소제곱법(weighted least squares method)을 통하여 평활의 정도를 조정가능

윈도우 : 산점도의 일부만 볼 수 있게 열어 둔 창문 틀의 크기

이동평균방법도 평활의 한 가지 방법임

산점도 평활법(LOWESS ; locally weighted regression scatterplot smoothing)에서 윈도우 너비를 너무 작게 잡으면 울퉁불퉁한 적합곡선으로 데이터를 표현하고, 윈도우 너비를 크게하면 데이터의 윤곽이 잘 살리지 못하는 밋밋한 형태의 곡선으로 데이터를 표현하게 됨