2강. 탐색적 자료분석[EDA] 시각화 I

◈ 담당교수 : 이태림

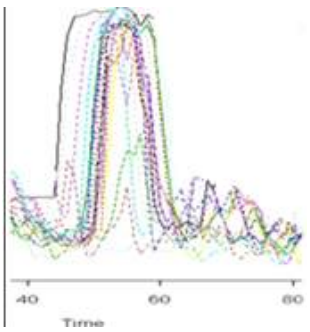
들어가기

■ 주요용어

용어	해설
탐색적 자료분석 (Exploratory Data Analysis)	데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석 기법
저항성 (Resesistance)	자료의 일부가 파손되었을 때 영향을 적게 받는 성질
재표현 (Re-expression)	원래 변수를 적당한 척도로 변환하여 분포의 대칭성, 분산 안정성, 선형성을 갖게 한다.
원그래프 (pie chart)	항목별 구성을 잘 나타내는 그래프
막대그래프 (bar graph)	항목별 발생 도수를 막대의 상대적인 길이로 나타내는 그래프

연습문제

1. 자료의 분포가 봉우리가 왼쪽으로 치우치고 오른쪽 긴 꼬리 분포를 가질 때 적용되어야 하는 EDA의 주제는?



① 자료의 재표현
② 잔차의 검토
③ 저항회귀식
④ 이원분석

정답 : ①

해설 : 자료의 분포가 대칭성을 갖지 못할 때 본격적인 분석에 앞서 변수를 재표현하

여 대칭성을 도모한다.

2. 일련의 면밀한 관찰 작업을 통하여 자료에 내재하는 구조적 관계를 알아내는 자료 분석 방법은?

① 잔차 분석(Residual Analysis)
② 확증적 자료 분석(CDA)
③ 통계그래프 분석(Statistical Graphics)
④ 탐색적 자료 분석(EDA)

정답 : ④
해설 : 탐색적 자료 분석은 데이터의 특징과 내재하는 구조적 관계를 알아내기 위한 기법들

3. 회귀모형을 추정하기 위해서 자료를 세군으로 나누고 요약점 (XL, YL), (XM, YM), (XR, YR)을 구하여 이상치에 영향을 받지 않는 기울기와 절편을 구하는 과정은 EDA의 추구하는 네 가지 주제 중 어떤 것에 적합한 예인가?

① 현시성
② 저항성
③ 잔차
④ 자료의 재표현

정답 : ②
해설 : 자료의 일부가 이상치가 포함되거나 자료의 일부가 파손되었을 때 영향을 적게 받는 성질

정리하기

1. EDA의 목적은 자료의 구조 및 특징을 파악하기 위하여 효과적이고 신뢰성 있는 자료의 요약과 그래프 기법의 활용에 있다.

2. 자료를 면밀히 검토해 보고자 하는 EDA입장에서의 데이터 시각화의 기본철학과 분석모형을 설명하고 탐색적 자료 분석의 성공 사례를 들어 EDA의 의의를 현실에서 확인해본다.

3. 저항성(resistance)
자료의 일부가 기존과 현격히 다른 값으로 대체되었을 때, 즉 자료의 일부가 파손

되었을 때 영향을 적게 받는 성질이다. 즉 저항성 있는 통계 또는 통계적 방법은 데이터의 부분적 변동에 민감하게 반응하지 않는다.

4. 잔차의 해석(residual)
잔차란 관찰값들이 주경향으로부터 얼마나 벗어났는지를 말해준다. 즉 잔차를 구해봄으로써 데이터의 보통과 다른 특징을 찾아내야 한다.

5. 자료의 재표현(re-expression)
자료의 재표현은 데이터분석과 해석을 단순화할 수 있도록 원래 변수를 적당한 척도(로그변환, 제곱근변환, 역수변환)로 바꾸는 것을 말한다. 이와 같은 변환을 통하여 분포의 대칭성, 분포의 선형성, 분산의 안정성, 관련변수의 가법성 등 데이터 구조파악과 해석에 도움을 얻는 경우가 많다.

6. 현시성(graphic representation)
현시성은 자료의 그래프에 의한 표현 즉 자료 안에 숨어있는 정보를 시각적으로 나타내줌으로써 자료의 구조를 효율적으로 잘 파악하게 된다는 것이다. 이런 의미에서 EDA에서는 다양한 그래프의 작성법들이 이용되고 있다.

7. 원그래프(Pie graph)
항목별 구성을 잘 나타내는 그래프.

8. 막대그래프(bar Chart)
항목별 발생 도수를 막대의 상대적인 길이로 나타내는 그래프.

참고자료

1. 허명회 (2014). 데이터 시각화, 자유아카데미.
2. 허명회 (2014). R을 활용한 탐색적 자료 분석, 자유아카데미.