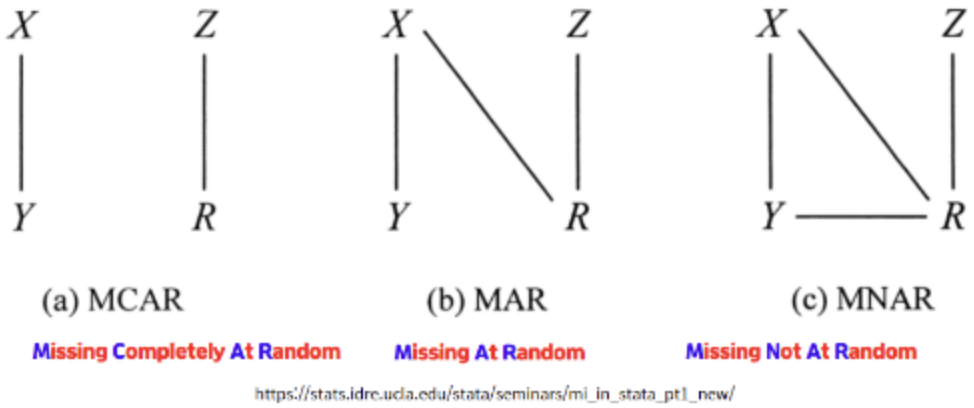
의미 및 종류

- 측정된 샘플에서 누락된 변숫값. 즉 존재하지 않은 데이터를 의미
- NA(Not Available)로 포현하지만 데이터를 수집하는 환경에 따라 null, 공백, -1 등 다양하게 표현될 수 있다.
- 결측치는 오류로 인해 발생할 수도 있지만, 조사 대상이 측정을 원하지 않을 때에도 발생(정치 성향같은 민감한 질문)
- 결측치는 원시 데이터에서 쉽게 찾아 볼 수 있는 오류로 이를 해결하기 위해서 결측치를 포함하고 있는 샘플을 삭제, 해당 변수 제거, 결측치 무시, 결측치 추정(평균, 증앙값 등의 통계량 or 머신러닝 기법을 활용하여 값 추정)

결측치 종류



MCAR

- 완전 무작위 결측(Missing Completely At Random)

- 순수하게 결측값이 무작위로 발생하는 경우
- 누락된 데이터의 이유는 나머지 데이터와 관련이 없다.
- 설문 조사에서 응답자가 실수로 질문을 놓친 예를 들 수 있다.
- 이런 경우는 결측값을 포함한 데이터를 제거해도 편향(Bias)가 거의 발생되지 않음

MAR

- 무작위 결측(Missing At Random)
- 다른 변수의 특성에 의해 해당 변수의 결측치가 체계적으로 발생한 경우

- 다른 열들의 데이터로부터 누락된 데이터를 유추 할 수 있다.
- 예를 들어, 특정 설문 조사 질문에 대한 응답 누락은 성별, 나이, 라이프스타일 등과 같은 다른 요인에 의해 조건부로 알아 낼 수 있다.

MNAR

- 비무작위 결측(Missing Not At Random)
- 결측값들이 해당 변수 자체의 특성을 갖고 있는 경우 곧 임의가 아닌 결측이다.
- 결측값에 대한 근본적인 이유가 있는 경우
- 예를 들어, 소득이 매우 높은 사람들은 그것의 공개를 주저하는 경향이 있다.

구조적으로 누락된 데이터

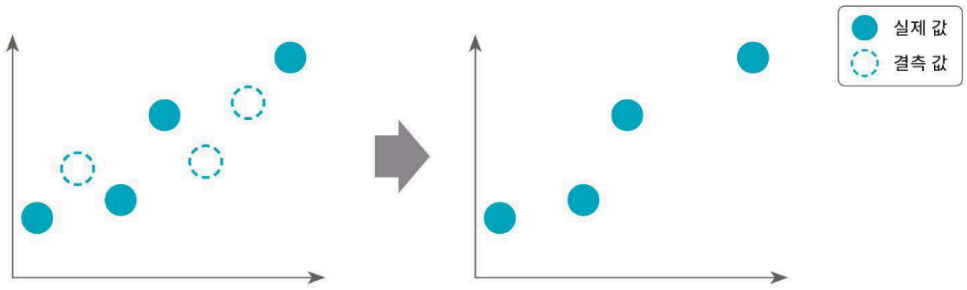
- 대개 MNAR 하위 집합인 경우가 많으므로 논리적 이유로 데이터가 누락된다.
- 예를 들어, 배우자의 나이를 나타내는 변수가 없으면 배우자가 없는 것으로 추정할 수 있다.

결측치 해결 방법

- 가장 간단한 결측값 처리 방법은 결측값이 많이 존재하는 변수를 제거하거나 결측값이 포함된 행을 제거하는 방법
- 전체 데이터에서 결측값 비율이 10% 미만일 경우 이 방법을 사용하는 경우가 많음
- 데이터의 갯수가 적을 경우 결측값을 삭제하면 데이터가 편중되어 편향이 발생할 위험도 존재

단순 대치법

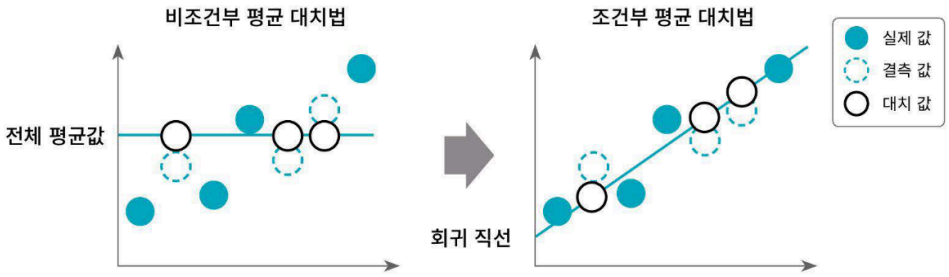
- 결측값이 존재하는 데이터를 삭제하는 방법
- 가장 쉬운 결측값 처리 방법이지만 결측값이 많은 경우 대량의 데이터 손실이 발생할 수 있음
- 데이터가 무작위로 누락되지 않은 경우 분석에서 이러한 관측값을 제거하면 결과에 편향이 생길 수 있다.
- 가장 쉬운 방법이지만 특히 작은 데이터셋의 경우 항상 좋은 방법은 아니다.



평균 대치법

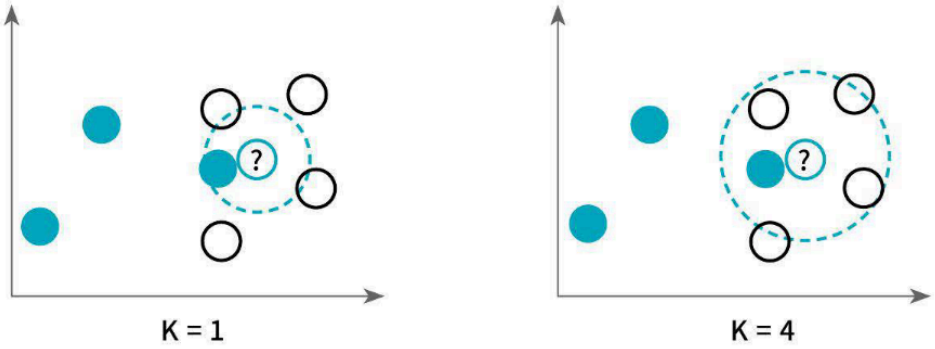
- 관측 또는 실험으로 얻은 데이터를 대표할 수 있는 평균 혹은 중앙값으로 결측값을 대치하여 불완전한 자료를 완전한 자료로 만드는 방법
- 범주형 변수의 경우는 최빈값을 사용할 수 있다.
- 비조건부 평균 대치법과 조건부 평균 대치법이 있다.
  - 비조건부 평균 대치법
    - 데이터의 평균값으로 결측값을 대치
  - 조건부 평균 대치법
    - 실제 값들을 분석하여 회귀분석을 활용한 대치 방법

- 단점은 데이터셋의 분산이 감소될 수 있다.



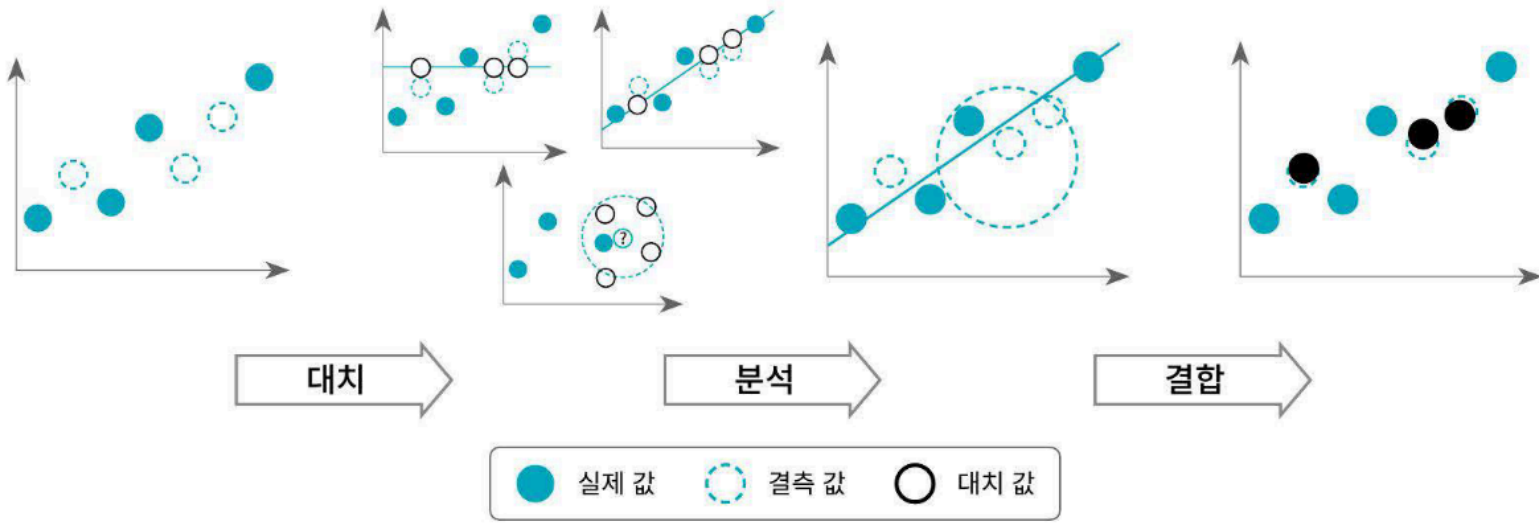
단순 확률 대치법

- 평균 대치법에서 추정량 표준 오차의 과소 추정 문제를 보완하고자 고안된 방법
- 대표적인 방법으로 K-Nearest Neighbor 방법이 있다.
- K-Nearest Neighbor 방법
  - K 최근접 이웃 알고리즘으로 주변 k개의 데이터 중 가장 많은 데이터로 대치하는 방법
  - k값의 값에 따라 결과가 달라지기 때문에 사용자가 선정해야 하지만 적절한 k값을 선정하기가 쉽지 않다



다중 대치법

- 여러 번의 대치를 통해 n개의 임의 완전자료를 만드는 방법
- 결측값 대치, 분석, 결합의 세 단계로 구성



클러스터링 결과 활용

- 신체 관련 지표의 경우 결측치를 정교하게 대체하려면 성별로 평균값이나 중앙값을 사용할 수 있다 .

보간법

- 데이터가 시계열적 특성을 가지고 있을 때는 보간법(interpolation)을 사용하는 것이 효과적임
- 매출 데이터의 일별 판매금액 변수의 결측값을 대치하고 하는 경우

단순 순서 보간법

기준일자	매출(만)
2022-04-17	2,800
2022-04-18	3,100
2022-04-19	3,000
2022-04-20	
2022-04-26	
2022-04-28	3,900
2022-04-29	4,200
2022-04-30	4,300

기준일자	매출(만)
2022-04-17	2,800
2022-04-18	3,100
2022-04-19	3,000
2022-04-20	**3,300**
2022-04-26	**3,600**
2022-04-28	3,900
2022-04-29	4,200
2022-04-30	4,300

시점 고려 보간법

기준일자	매출(만)
2022-04-17	2,800
2022-04-18	3,100
2022-04-19	3,000
2022-04-20	
2022-04-26	
2022-04-28	3,900
2022-04-29	4,200
2022-04-30	4,300

기준일자	매출(만)
2022-04-17	2,800
2022-04-18	3,100
2022-04-19	3,000
2022-04-20	**3,100**
2022-04-26	**3,700**
2022-04-28	3,900
2022-04-29	4,200
2022-04-30	4,300

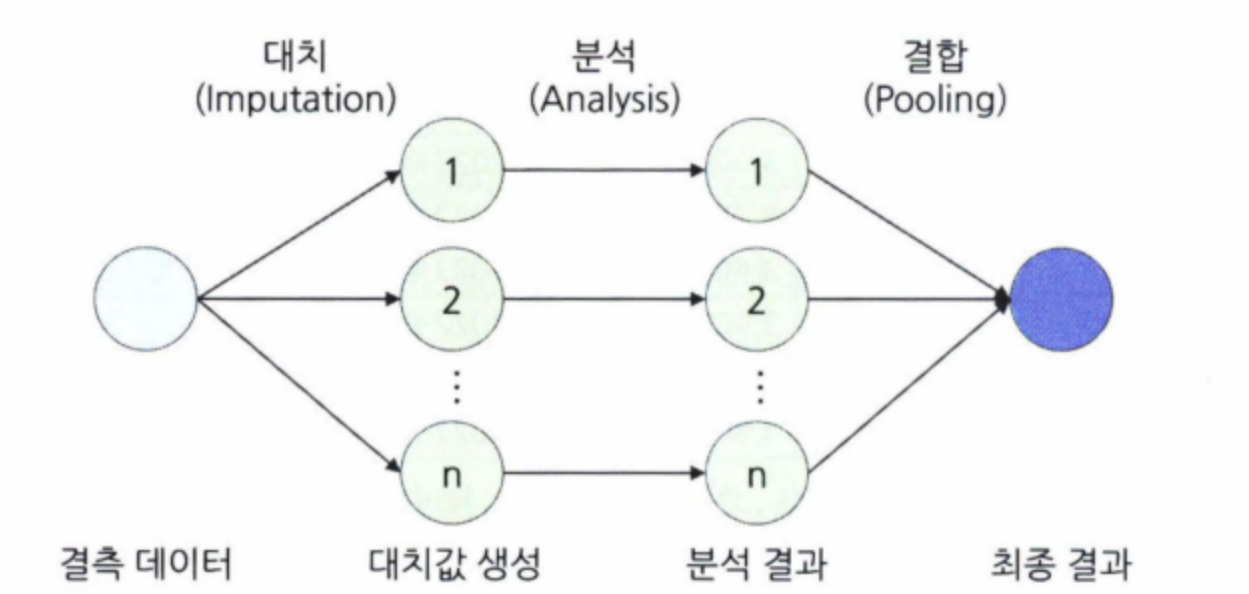
- 위의 그림에서 4월 20일의 판매금액은 19일이나 21일의 판매금액과 비슷할 것으로 기대할 수 있음
- 그렇기 때문에 전 시점 혹은 다음 시점의 값으로 대치하거나 전 시점과 다음 시점의 평균 값으로 대치하는 방법을 사용하는 것
- 다만 시점 인덱스의 간격이 불규직하거나 결측값이 두 번 이상 연달아 있을 때는 선형적인 수치 값을 계산해 보간하는 방법을 사용

회귀 대치법

- 회귀 대치법(regression imputation)은 회귀식을 이용하여 결측값을 추정
- 단순하게 평균값 등을 대치하는 것아 아니라 해당 변수와 다른 변수 사이의 관계성을 고려하여 결측값을 계산하면 보다 합리적으로 결측값을 처리할 수 있음
- 예를 들어 연령 변수의 결측값을 대치하기 위해 '연 수입' 변수를 사용하는 것
- 만약 데이터가 수입이 많아질수록 연령이 높아지는 상관관계를 가지고 있다면, 해당 관측치가 연 수입이 높으면 상대적으로 높은 연령을 추정하여 결측값을 대치하는 것
- **추정하고자 하는 결측값을 가진 변수를 종속변수로 하고, 나머지 변수를 독립변수로 하여 추정한 회귀식을 통해 결측치를 대치**
- 이 경우에도 결측된 변수의 분산을 과소 추정하는 문제를 가지고 있어, 해결하기 위해 인위적으로 회귀식에 확률 오차항을 추가하는 확률적 회귀대치법(stochastic regression imputation) 을 사용하여 변동성을 조정
  - 관측된 값들을 변동성만큼 결측값에도 같은 변동성을 추가해 주는 것
  - 하지만 이 방법도 여전히 어느 정도 표본오차를 과소 추정하는 문제를 가지고 있음

다중 대치법

- 다중 대치법(multiple imputation)은 단순 대치법들의 표본오차 과소 추정 문제를 해결하기 위해 많이 사용되는 방법이다.
- 단순대치를 여러 번 수행하여 n개의 가상적 데이터를 생성하여 이들의 평균으로 결측값을 대치하는 방법으로 다음 3가지 단계로 구분할 수 있다. - 대치 단계(Imputations step) - 가능한 대치 값의 분포에서 추출된 서로 다른 값으로 결측치를 처리한 n 개의 데이터셋 생성 - 분석 단계(Analysis step) - 생성된 각각의 데이터셋을 분석하여 모수의 추정치와 표준오차 계산 - 결합 단계(Pooling step) - 계산된 각 데이터셋의 추정치와 표준오차를 결합하여 최종 결측 대치값 산출



- 대치 단계에서는 일반적으로 몬테카를로(MCMC: Markov Chain Monte Carlo) 방법이나 연쇄방정식을 통한 다중 대치(MICE: Multivariate Imputation by Chained Equation)를 사용하여 대치값을 임의로 생성한다.
- 가상적 데이터는 너무 많이 생성할 필요는 없고 5개 내외 정도만 생성해도 성능에 큰 문제가 없음
- 다만 결측값의 비율이 증가할수록 가상데이터도 많이 생성해야 검정력이 증가
- 만약 몬테카를로 방법을 사용하여 5개의 데이터셋을 생성했다면, 각 데이터셋의 결측값들은 난수로 생성됐기 때문에 모두 다를 것
- 평균공식을 통해 각각의 데이터셋의 상이한 추정치와 표준오차를 결합하여 결측치가 채워진 최종 데이터셋을 만들면 다중 대치법의 모든 단계가 완료된다.