**ㄱ**

**- 검정통계량**

: Z값(표준화) 혹은 t값(스튜던트화)를 주로 사용하며, 검정통계량의 분포는 귀무가설이 참이라는 가정 아래 표준정규분포나 t값이 된다. P값은 검정통계량 값보다 귀무가설을 기각하는 방향으로 더 심하게 검정통계량 값이 관측될 확률을 의미한다.(검정통계량 분포에서의 확률). F값은 MSR/MSE 이며 회귀모형의 유효성을 의미한다. 3.5 이상이면 유의미한 모형으로 판단한다.

**- 공분산/상관계수**

: Cov(X,Y) = E[(X-E(X))(Y-E(Y))]. 오직 선형관계만 파악할 수 있으면 그 크기에 따른 강도 파악은 어렵다. 따라서 공분산을 각각 X와 Y의 표준편차를 곱한 값으로 나눈 상관계수를 사용한다. 상관계수 또한 선형관계와 그 강도만 파악할 수 있기 때문에 독립성이나 곡선관계는 알 수 없다.

**- 공분산분석/ANCOVA**

: 다중 회귀분석 모형과 비슷하지만, 관심대상은 명목형 변수이다. 이때 혼란변수인 메트릭 독립변수를 공변량이라하고, 이 공변량에 의한 효과를 제거하기 위해 다중회귀분석 후 분삭분석으로 명목척도의 분석이 이루어진다. 메트릭변수가 유의함이 밝혀지면 그 변수를 평균값 등으로 통제후 관심변수의 추정값을 확인한다.

**- 교차분석/카이제곱 검정(범주형 자료 분석)**

: 일반적으로 빈도나 분포에 대해 가설검증을 하거나 범주형 변수끼리의 독립성을 검증하는 데 사용된다. 카이제곱 값은 편차의 제곱값을 기대빈도로 나눈 값들의 합이다. 이 때 자유도는 (r-1)(c-1)이다. 적합도 검정(분포가 고른지), 독립성 검정(가설검정), 동질성 검정(편향된 데이터인 경우) 3가지로 분류 가능하다.

**ㄷ**

**- 다차원 척도법/MDS**

: 개체들 사이의 유사성/비유사성을 측정하여 공간상에 점으로 표현해 집단을 시각적으로 표현하는 분석방법. 유클리드 거리행렬을 이용하며, 스트레스 값이 0에 가까울수록 적합도 수준이 완벽하고, 1에 가까울수록 나쁘다.

**- 등분산성**

: 비교집단끼리의 분산이 같다는 가정. 분산안정화를 위해서는 루트, 로그 변수변환을 사용하여 잔차를 정규분포로 만들어준다.

**ㄹ**

**- 랜덤 포레스트(앙상블)**

: 의사결정나무는 분산이 크므로 배깅과 부스팅보다 더 많은 무작위성을 주어 약분류기를 생성 후 이를 선형결합하여 최종 학습기를 만드는 방법이다. 부트스트랩 데이터를 추출하고, 이들을 훈련 후 대표 변수 샘플을 도출하여 리프노드로 분류한다. 그리고 이 노드들의 선형 결합으로 최종 모델을 결정한다. 트리의 개수, 최대 깊이, 임의성 이 주요 파라미터이다.

**- 랜덤화/랜덤 샘플링**

: 랜덤화란 표본을 실험그룹과 대조그룹으로 할당하는 과정(내적타당성)이며, 랜덤 샘플링은 모집단에서 표본을 선택하는 과정(외적 타당성)이다.

**- 로지스틱 회귀모형**

: 로그오즈 . 비선형 회귀모형이며, 로지스틱 함수를 사용하는 로짓 모형과 누적분포함수를 이용하는 프로빗모형이 있다. 최대우도법을 사용하여 추정한다.

**ㅂ**

**- 배깅(앙상블)**

: 학습 데이터에서 다수의 부트스트랩 자료를 생성하고 각각의 자료에 대해 모델을 생성하여 최종 다수결로 분류하는 기법

**- 범주형 자료 분석**

: 분할표 분석, 교차분석(카이제곱 검정), 피셔의 정확검정 등이 있다.

**- 베이즈 정리**

: 사전확률과 사후확률 사이의 관계를 이용하여 이 중 하나를 알 때 반대도 알 수 있다. 조건부확률과 전 확률의 정리를 활용한다.

**- 부스팅(앙상블)**

: 잘못 분류된 개체들에 가중치를 적용, 약분류기를 결합해 강분류기를 만든다. 동일 가중치 분류기로 시작해서 가중치 변경을 이어간다.

**- 부트스트랩(배깅)**

: 동일한 크기의 표본을 랜덤복원추출로 뽑은 자료. 100개의 샘플을 추출하더라도 한번도 선택되지 않는 데이터가 36.8%가 된다.

**- 분산분석/ANOVA**

: 두 개 이상의 집단 간 비교를 할 때 집단 간 분산 비교로 얻은 F-분포를 이용하여 가설검정을 수행하는 방법. 모든 그룹의 평균이 같다는 귀무가설에서 시작한다. SST(총제곱합) = SSB(그룹간 제곱합) + SSE(그룹내 제곱합). F값(MSB/MSE)으로 검정하며, 자유도는 rc-1 = (c-1) +c(r-1) 이다(r : 한 그룹 내 관측값 수, c: 그룹수). 2개 요인 이상 분산분석 시 상호작용효과가 존재하는지 유의하며, 이 경우 그룹을 나눠서 비교해야한다.

**- 분할표 분석(범주형 자료 분석)**

: 일반적으로 행은 독립변수, 열을 종속변수로 배치한다. 특히 2원 분할표를 많이 사용. 상대위험도, 오즈비로 활용한다.

**- 블록화**

: 혼란변수 통제를 위해 비슷한 특성을 가진 블록으로 구분하며, 그룹내 분산을 줄일 수 있다.

**- 분산확대인자/VIF**

: 로 나타낸다. 5 이상이면 문제가 있다고 의심하며, 10 이상이면 높은 다중공선성으로 인한 문제가 발생할 수 있다고 판단한다.

**ㅅ**

**- 서포트 벡터 머신/SVM**

: 데이터를 분리하는 초평면 중 데이터들과 가장 먼 초평면을 선택하여 분리하는 이진선형분류 모델이다. 변수의 모든 속성을 활용하며 과대적합의 가능성이 낮다. 마진의 오분류를 허용하느냐의 여부로 하드/소프트 마진 SVM으로 나뉜다. 선형으로 분리가능한지로도 나뉜다.

**- 신뢰성/타당성**

: 측정에서의 신뢰성이란 반복된 측정이 어느 정도의 일관성을 나타내는가 이다. 상호교환적인 두 측정의 상관계수를 일반적으로 신뢰성으로 본다. 타당성이란 관심 있는 개념을 측정값이 얼마나 잘 나타내고 있는지를 의미한다.

**- 심슨의 패러독스**

: 다른 원인변수에 의해 결과가 왜곡되는 현상. 어떠한 수준에 대해 한 그룹이 많이 포함되어 있다면(ex. 소득이 높은 그룹에 남자가 많이 할당되어 소득이 높은/낮은 두 그룹에선 모두 남성의 상품 구매율이 낮으나 전체 자료에선 남성의 구매율이 여성보다 높은 경우) 전체 결과가 왜곡될 수 있다. 가중치 보정하거나 적절한 샘플링을 통해 해결해야 된다.

**- 쌍체 T검정**

: 단일그룹 사전사후측정으로, [O1 X 02] 에서 O2 – O1을 행하여 처리효과를 측정하는 검정 방법이다. 반복측정으로 확장될 수 있다. 혹은 두 대상이 의미있게 다른지를 검정하는 과정이다.

**ㅇ**

**- 아리마 모델/자기회귀누적이동평균모형**

: 비정상 시계열 모형으로, AR이나 MA, ARMA 등으로 정상화할 수 있다. p(AR모형), q(MA모형), d(ARMA로 정상화할 때 몇 번 차분했는지)라는 차수를 가지고 있다.

**- 앙상블**

: 여러가치 모형들의 예측/분류 결과를 종합하여 최종적 의사결정에 활용하는 기법. sampling(배깅, 부스팅)과 variable(랜덤 포레스트)로 나눌 수 있다.

**- 연관성 분석**

: 항목 간의 상호관계 혹은 종속 관계를 찾아내는 분석기법. 지지도(교집합), 신뢰도(조건부확률), 향상도(신뢰도/P(B)) 등을 활용하여 나타낸다.

**- 오차**

: 단순선형 회귀 모형에서의 오차는 어떤 확률에 의해 랜덤으로 나타나는 변수이다. 이로 인해 Yi도 랜덤 변수가 되어 E(Yi) = α + βi\*xi, Var(Yi) = 가 된다.

**ㅈ**

**- GBM(부스팅)**

: gradient descent를 이용해 학습시키는 부스팅 기법. 이때 negative gradient는 residual이 된다. 여기에 CART 알고리즘을 더하면 XGBM이 되는데, CART 란 지니 계수를 불순도 척도로 이용하여 지니계수가 가장 낮은 조건조합을 찾아 분기하는 알고리즘이다. LightGBM은 기존 트리부스팅 기법들이 level-wise(BFS)인 반면 leaf-wise(DFS)로 확장시키면서 속도가 빠르고 큰 데이터일수록 정확해진다. CatBoost는 random permutation으로 ordered boosting을 진행하고, 범주형 데이터에 강한 ordered target encoding을 활용한다.

**-주성분 분석**

: 상관관계가 있는 고차원 자료를 최대한 보존하면서 저차원 자료로 변환하는 차원축소 방법. 기존의 상관성이 높은 변수들을 요약, 축소한다. 고윳값이 높은 순서대로 고유벡터만으로 데이터를 복원한다.

**ㅊ**

**- 체계적 오차(신뢰성/타당성)**

: 측정과정에서 일정한 패턴을 갖는 체계적 바이어스를 갖는 오차. 타당성과 반비례 관계를 가지고 있다.

**- 최대우도법(로지스틱 회귀모형)**

: likelihood, 즉, “자료가 추정된 분포로부터 나왔을 가능성”을 최대로 하도록 모형을 맞추는 방법이다. 즉, 이미 뽑은 표본이 발생할 확률을 최대로 만드는 모수의 추정값(여기선 베타값)을 찾는 과정이다.

**ㅎ**

**- 회귀모형의 해석**

: SST(총변동) = SSR(회귀제곱합) + SSE(오차제곱합). 자유도는 SSE = n-k(베타 개수)- 1(인터셉트), SSR = k, SST = n-1(평균을 사용했으므로) 가 된다. 최소제곱법(L2)으로 a,b를 찾는다.