In [ ]:

04.03 스케일링

- OLS모형으로 회귀분석 실시 -> 조건수 큰 경우를 해결하기 위한 방안 1)스케일링

1) 조건수
    - 가장 큰 고유치와 가장 작은 고유치의 비율
    - 회귀분석에선, 공분산행렬의 가장 큰 고유치와 가장 작은 고유치 비율
    - 회귀분석에선, 조건수는 항상 양수 (공분산 행렬 = 대칭 + 양의정부호행렬. 양의정부호 = 고유값 모두 양수)
        *공분산행렬은 대칭 + 양의정부호! (그래야 역행렬 가능)

    - 조건수가 크면, 발생하는 오차에 대한 solution의 민감도가 커진다. (4page)

    - 조건수가 가장 작은 경우(단위행렬, 조건수 =1), 가장 큰 경우(힐버트행렬, 조건수 = 15,000) (4, 5page)

    - 결국, 회귀분석 시, 공분산행렬의 조건수가 크면 회귀분석을 사용한 예측값도 오차가 커진다.

      조금의 오차(계수행렬A, 상수벡터에b)가 있어도
      solution set(x)가 매우 크게 달라져 제대로된 회귀분석, 예측을 할 수 없다.

2) 회귀분석과 조건수
    - 회귀분석에서 조건수가 커지는 경우 2가지 (6page)
        1. 변수들의 단위 차이(m와 mm) -> 스케일이 크게 달라짐 -> 조건수 커짐 <== 스케일링(평균, 표편일정하게)
        2. 다중공선성 -> 조건수 커짐 -> 변수 선택, PCA를 통한 차원축소(고유값 큰것들 위주로 남기기)

In [ ]:

04.04 범주형 독립변수

- OLS모형으로 회귀분석 실시 -> 조건수 큰 경우를 해결하기 위한 방안 1)스케일링 -> 2)범주형 독립변수 다루기

- 범주형 독립변수를 갖는 경우의 회귀모형 -> 더미변수화! (원핫인코딩)
  ex) 혈액형 4개인 경우, (1,0) 이 아니라, (1,0,0,0), (0,1,0,0) 이렇게 가줘야 함!
  why? 분석 결과 해석을 하려면 특징 갯수만큼 원소를 갖는 카테고리 확률변수로 표시해줘야 함

- 더미변수화 : 풀랭크 방식(상수항 X), 축소랭크 방식(상수항 O = 기준값 존재)

1) 풀랭크 방식
   - 기준 없이, 각 독립변수의 주체적인 변화를 분석하고 싶다면 풀랭크!
   - 더미변수의 가중치는 각각 상수항이 됨 = y절편 (1-3page)

2) 풀랭크 방식 OLS 예시 (5,7,8,9)

3) 축소랭크 방식
   - 기준을 두고, 그 기준 대비 각 변수들의 변화를 분석하고 싶다면 축소랭크!
   - 예) H0 : 1월기온 = 2월기온 <==> H0 : w2 = 0 (기준 1월, 축소랭크 OLS)
   - 더미변수의 가중치 = 기준값의 가중치 + 추가적으로 더해지는 가중치 (3page)

4) 축소랭크 방식 OLS 예시 (10,11,12)

5) 보스턴 집값 데이터의 범주형 변수 (13page)

6) 두 개 이상의 범주형 변수가 있는 경우 (19page)
- 범주형 변수가 2개 이상이면 통합축소형 **or** 상호작용방식 (A,B / X,Y / 등..)

7) 범주형 독립변수와 실수 독립변수의 상호작용
- 실수독립변수에 영향을 주는 범주형 독립변수가 있다면, 상호작용항 생성해줘야함 (21page)

In [ ]:

04.05 부분회귀

- Q : 기존의 모델에 새로운 독립변수를 추가하면, 가중치의 값이 달라질까? 그렇다.
    con) 종속변수에 영향을 미치는 모든 독립변수를 회귀모형에 포함하지 않는 한 모형의 가중치는 항상 편향된 값이다.

con)
    1. 처음에는, 독립변수를 최대한 많이 넣고 가중치 학습시키는 것이 좋음
       (새로운 변수 추가 시, 가중치 값이 달라짐)
    2. 순수한 상관관계를 확인하고 싶다면, 부분회귀로 각 변수의 순수한 스캐터플롯을 그려야 함
       (그냥 스캐터플롯으론 다른 변수들의 영향이 복합되어있어 정확한 상관관계를 확인하기 어렵다.)

1) 새로운 독립변수가 추가된다면, 가중치는?
    변한다. 안변하는 경우는 1) 새 변수와 종속변수가 독립일 때 or 2) 새 변수와 기존 변수들 간 독립일 때
    증명해보기, 프리슈-워-로벨 정리

2) 부분회귀 플롯
    - 각 변수들의 종속변수와의 순수한 상관관계를 볼 수 있음 (3-4p)

3) CCPR 플롯
    - 부분회귀 플롯과 비슷한 역할. 하지만, 조금 정석적인 방법은 아님

4) plot_regress_exog : 부분회귀 플롯, CCPR플롯 함께 보여줌 (12page)

In [ ]:

05.01 확률론적 선형 회귀모형

- 예측에는 오차가 있다. 그런데 그 오차가 얼마인지 모른다면? 아무것도 모르는 것과 같다.
- 오차가 어느정도 나는 건지 알아야 한다 => 확률론적 선형 회귀모형

1) 방법 1 : 부트스트래핑 (1page 상단)
    - 표본 데이터가 달라질 때, 회귀분석의 결과는 어느정도 영향을 받는지를 알기 위한 방법
    - 기존의 데이터를 re-sampling(재표본화) -> 회귀분석 재실시 (w추정)
    (기존의 N개 데이터에서 N개 데이터 선택하되, 중복 선택 가능하게 함)
    - 이 방법은 연산소요시간이 너무 오래걸림 (1000번 리샘플링 - 회귀분석 반복!)

2) 방법 2 : 확률론적 선형회귀모형 (부트스트래핑 대체 방법)
    - 사실 OLS 모델 자체는 확률과 상관 없는 모델 ==>> 여기에 확률론을 접목시킨 것
    - 데이터 = 확률변수에서 생성된 표본이라 가정
    - 4가지 가정을 세팅한 것이 기본모형 => 확률론적 선형회귀모형
        1) 선형 정규분포 가정
        2) 외생성 가정 (exogeneity)
        3) 조건부 독립 가정
        4) 등분산성 가정

    1) 선형 정규분포 가정 : 종속변수 y는 정규분포를 따르고, 기대값(모수)은 독립변수 x의 선형조합으로 결정됨(5p)
        <==> 잡음(y-기대값)은 0주변에서 분포한다.
        <==> 잡음이 정규분포일 뿐이지, x나 y 주변확률분포 자체가 정규분포일 필요 없음

    2) 외생성 가정 : 잡음의 기대값은 x와 상관없이 0이다. (5p)
        <==> 따라서, 잡음의 무조건부 기댓값 = 0, 잡음과 독립변수 x는 독립 임을 증명할 수 있음

    3) 조건부 독립 가정 : i번째 데이터에 대한 잡음, j번째 데이터가 갖는 잡음은 서로 독립(공분산 = 0) (5p)
        <==> E[잡음i*잡음j] = 0
        <==> 잡음벡터의 공분산행렬 = 대각행렬(비대각성분 = 0) (왜냐면 cov(잡음i,잡음j)=0)

    4) 등분산성 가정 : 데이터에 관계없이 잡음의 분산 값은 일정하다 (5p)
        <==> 3) 조건과 함께, 결국 잡음벡터의 공분산 행렬은 항등행렬(대각성분 = 분산, 모두 같은 값)

3) 확률론적 선형회귀 모형 : 최대가능도 방법을 사용한 선형 회귀분석(OLS 방법처럼 찾을 수 없음)
        - but, 결과는 OLS와 확률론적 모형(MLE)는 같다.

- 차이점 : MLE는 정답(w의 정답)이 존재한다는 가정 하, 정답과 가장 유사한 것을 찾아보자라는 접근(하지만 정답 근처의 다른 답이 나옴)
  OLS는 정답 없이 그냥 RSS를 최소화 하는 값 찾은 것

4) 잔차의 분포
- 확률론적 선형회귀모형 "잔차 = e = y-w.Tx 도 정규분포따름"
- 확률론적 선형회귀모형에서는 잔차와 잡음이 다른 개념이다. (7page 상단)