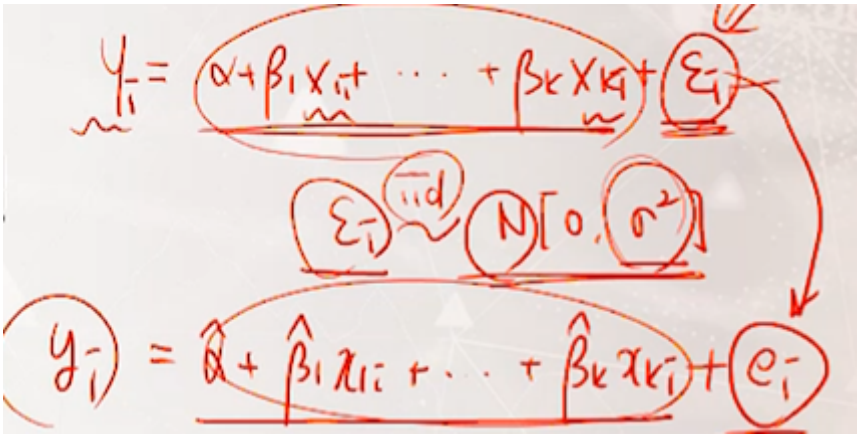
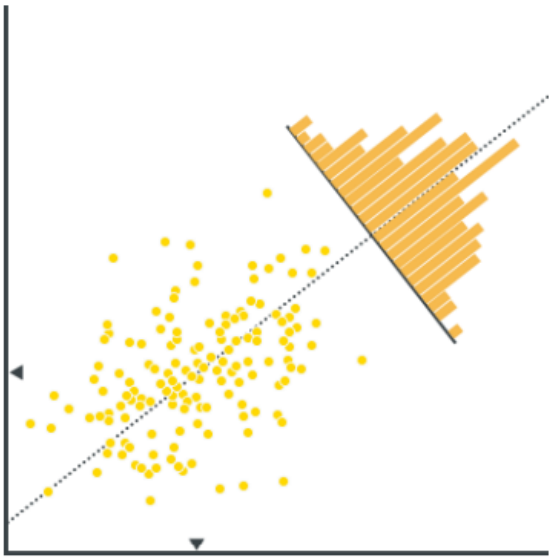
13.다중회귀분석(잔차분석, 다중공선성)

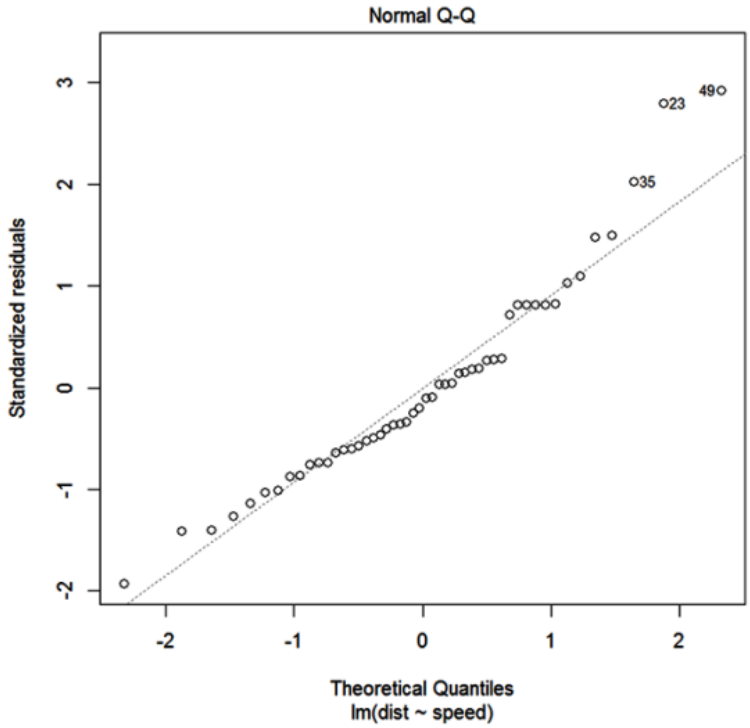
- 정리내용
  - 오차에 대한 분포 가정 : 정규, 등분산, 독립
  - 오차에 대한 가정이 있어야 검정이 가능
  - 회귀분석의 마지막 단계로 오차에 대한 가정이 적절한지 여부 확인
  - 오차에 대한 추정치 개념인 잔차를 이용하여 분석(잔차분석)
  - 독립변수들 간 강한 상관관계가 부정적 영향을 미치는 현상(다중공선성)
- 가정 위반 검토및 해결
  - 다중회귀모형의 가정 위반 검토 및 해결
    - 잔차분석
      - 회귀 모형에서의 가정이 적절한 것인가에 대한 평가



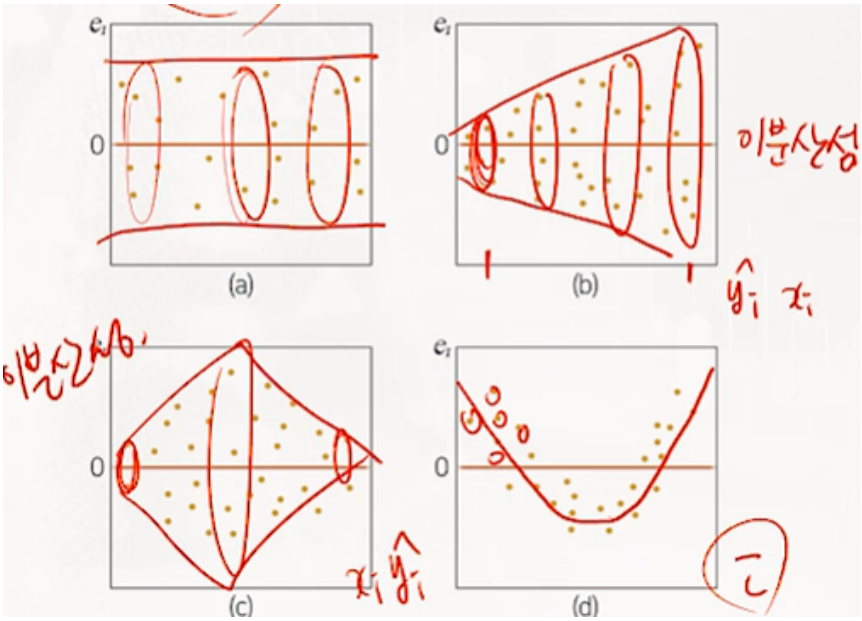
  - 1) 오차의 정규성 2) 오차의 등분산성 3) 오차의 독립성
  - 오차는 확률변수로 관찰되지 않는 값이므로, 각 오차들에 대응되는 잔차를 관찰한 뒤 잔차들의 분포를 통해 오차에 대한 가정의 적정성을 확인 할 수 있음



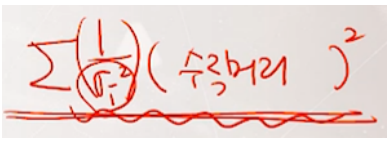
- 잔차 분석 방법
  - 각 가정 별로, 검정을 통한 방법과 그래프를 통한 시각적인 확인 방법이 가능
  - 시각적 방법을 이용할 경우,
    - 오차의 정규성 위반 : 히스토그램, QQ플롯
      - n개의 잔차에 따라 히스토그램을 그림
      - 히스토그램의 구간의 간격에 따라 fluctuation이 많이 달라짐
        fluctuation : (방향,위치,상황의)변동, 오르내림, 파동
      - 어떤 특정분포에 따르는가에 관한 그래프는 QQ플롯이 선호됨



- 오차의 등분산성 : 잔차산점도
- 오차의 독립성 : 잔차산점도
- 그래프 해석
  - 최소제곱법에 의한 성질로 인하여 항상 평균이 0에 수렴함
  - (a) : 등분한 가정을 만족, 특별한 함수가 보이지 않으므로, 독립가정 만족
  - (b) : x축이 변화함에 따라 분포도 달라지며, 이는 이분산성을 나타냄
  - (c) : 이 역시 분산에 문제가 있는 이분산성을 나타냄
  - (d) :
    - (1) x축이 시간에 흐름을 나타낼때 :
      대표적인 독립성 가정이 깨진경우의 그래프의 특징을 나타냄
    - (2) x축이 x_i일 경우 :
      x와 y가 애초에 선형이 아니였을 가능성이 농후함



- 가정 위반 시 해결방안
  - 오차의 정규성 위반 : 변수변환
    - 로그변환이나 제곱근변환을 통하여 정규성으로 변환 가능
  - 오차의 등분산성 : 가중최소제곱회귀
    - 각각을 각 자료값이 가지는 분산의 역수로 가중치를 두고 가중최소제곱을 사용
    - 가중최소제곱을 사용 시 우리가 얻는 효과
      - 변동성이 큰 자료는 적은 비중으로 그 계수 추정을하는데 반영
      - 변동성이 작은 자료는 그만큼 안정적인 자료라는 것을 의미



  - 오차의 독립성 : 시계열 분석
    - correlation의 구조가 어떠한지 우선 판단
    - 정상성 조건을 만족하는지 판단 후 만족한다면 어떤모델을 적용하면 좋을지 그 Autocorrelation의 구조를 먼저 파악한 다

음 적절한 모델을 선택해서 분석하는 것이 좋음

- 다중공선성
  - 다중공선성이란
    - 독립변수들 간에 강한 선형관계가 존재하는 경우
    - 다중회귀모형 분석 시 자주 발생하는 문제 중 하나임
    - 다중회귀모형에서 **회귀계수 추정**에 부정적인 영향을 미침
      - 1) 개별적인 회귀계수 추정의 신뢰성이 떨어져 추정치를 믿을 수 없음
      - 2) 전박적인 모형의 적합성이나 정확도는 크게 변하지 않음

  - 다중공선성 진단방법
    - VIF 계수 도출 (VIF : Variance Inflation Factor)
      $VIF = \frac{1}{1-r_j^2}$

      1) R_j^2 : x_j 종속변수로 두고 나머지 독립변수로 설명하는 다중선형회귀모델에서의 결정계수
    - VIF계수가 5 또는 10이상인 경우 다중 공선성이 심각한 것으로 판단
  - 다중공선성의 해결책
    - 변수선택으로 중복된 변수를 제거
    - 주성분 분석등을 이용하여 중복된 변수를 변환하여 새로운 변수 생성

- 릿지, 라쏘 등으로 중복된 변수의 영향력을 일부만 사용
-