Chapter 04. 회귀와 예측 - 보아

▼ 4.1 단순선형회귀

1. 회귀식
   - 독립변수 X → 종속변수 Y 관계를 식으로 나타낸 것
   - $Y = b_0 + b_1 X$

b_0	절편, 인터셉트(intercept)
b_1	기울기, 계수
Y	응답변수, 종속변수, 목표벡터
X	독립변수, 예측변수, 피처벡터

2. 적합값과 잔차
   - 모든 데이터가 회귀선 안에 들어오지 않음 → 오차항(e) 포함

     → $Y_i = b_0 + b_1 X_i + e_i$
   - 적합값 = 예측값 ($\hat{Y}_i$)

     → $\hat{Y}_i = \hat{b_0} + \hat{b_1} X_i$
   - 잔차 = 예측 오차 ($\hat{e}_i$ 원래 값에서 예측한 값을 빼서 구함)

     → $\hat{e}_i = Y_i - \hat{Y}_i$
3. 최소제곱
   - 회귀선은 잔차 제곱(RSS)를 최소화하는 선
   - 잔차제곱합을 최소화하는 방법 : **최소제곱회귀** or **보통최소제곱**(ordinary least squares **OLS**)

     $RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{b_0} - \hat{b_1} X_i)^2$

     → 추정치 $\hat{b_0}$ 과 $\hat{b_1}$ 이 RSS를 최소화하는 값임.
     - $\hat{b_1} = \frac{S_{xy}}{S_{xx}}$ (위의 값을 편미분하여 계산)
     - $S_{xy} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$,
     - $S_{xx} = \sum(X_i - \bar{X})^2$
     - $\hat{b_0}$ =$\hat{Y}_i - \hat{b_1} X_i$

- 특잇값에 민감함

4. 예측 대 설명
   - 전통적으로 $X \rightarrow Y$ 의 관계를 밝히고 회귀방정식의 기울기(회귀계수)를 구하는 것이 목표였음

     : 현상을 설명하는 데에 초점!

     ex) 교육 수준이 경제적 불평등(소득 수준)에 미치는 영향
   - 현재는 $\hat{Y}$의 값을 예측하는 모델을 구성하는데 사용됨

     : 예측에 더 초점!

     ex) 마케팅 비용 증가에 따른 매출 예측

▼ 4.2 다중선형회귀

예측 변수가 여러개 회귀식 → $Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e$

입력 변수가 여러개 일 때 최소제곱법으로 무리가 있음 → 모형 평가를 통해 오차를 최소화하는 최적선을 그려야함

1. 모형 평가
   - 제곱근평균제곱오차(RMSE)
     - $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$
   - 잔차 표준오차(RSE)
     - 독립 변수가 p개일 때

       → $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-p-1)}}$
     - 제곱근평균제곱오차와의 차이는 분모가 데이터 수가 아닌 자유도라는 점임

       ▼ 참고

       분모에 자유도를 사용하는 이유 : 표본을 통해 모집단의 분산을 구할 시 분모에 자유도를 사용하면 추정값에 편향이 발생하지 않음
   - R^2 : 결정계수
     - 0~1의 값을 가지며 모델 데이터의 변동률을 측정함. 1에 가까울수록 설명력이 높음.
     - $1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} = \frac{\sum (\hat{y}_i - \bar{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$
     - 수정 R 제곱 *adjusted R-squared*

        - 예측변수가 많을수록 R^2의 값이 올라감
        - 수정 R^2은 더 많은 예측변수를 추가하는 것에 대해 페널티를 가함
        - 다중회귀분석에서는 큰 차이 없음

- t 통계량 & p-value
    - t 통계량이 높을수록, p-value가 낮을수록 예측변수는 유의미함(*별이 뜨는 변수!)
    - 사고절약의 원리(Principle of Parsimony)를 위해 유의미한 예측변수를 고를 때 유용하게 활용할 수 있음.

2. 교차타당성검사(교차검증)
    - 1의 지표는 모두 표본 내 지표 = 모델을 구하는 데 사용했던 데이터를 똑같이 그대로 사용
    - 빅데이터의 출현으로 인해 표본 밖 유효성 검사가 유의미해짐
        - 홀드아웃 (train, test set 분리)
        - k-fold CV

          1) 1/k의 데이터 홀드아웃 샘플로 분리

          2) 남아 있는 데이터로 훈련 이후 1)을 이용해 모델 평가

          3) 데이터의 첫번째 1/k를 복원하고 다음 1/k(앞서 선택했던 레코드 제외)을 따로 보관

          4) 1)~3)을 반복

          5) 모든 레코드가 홀드아웃 샘플로 사용할 때까지 반복

          6) 모델 평가 지표를 평균과 같은 방식으로 결합

3. 모형 선택 및 단계적 회귀
    - 오컴의 면도날 : 모든 것이 동일한 조건에서 복잡한 모델보다는 단순한 모델을 우선 사용해야함
    - 변수가 추가될수록 rmse는 항상 작아지고 r2은 항상 커짐 → 수정 R 제곱을 사용
    - AIC 값을 최소화하는 모델
        - $AIC = 2P + n\log\left(RSS/n\right)$
        - P 변수의 개수, n 레코드의 개수. 변수를 k개 추가하면 2k만큼 불이익을 받음
    - AIC를 최소로 하거나 수정 R 제곱을 최대로하는 모델 찾는 방법

    - 부분집합회귀 : 모든 가능한 모델 검색
    - 단계적 회귀
        - 후진제거 : 별로 의미없는 변수를 연속적으로 삭제
        - 전진선택 : 상수 모델에서 시작하여 연속적으로 변수 추가
        - 단계적 회귀분석 : 전진 + 후진. 전진 선택의 방법으로 기여도가 높은 독립 변수부터 추가 후 변수를 단계별로 검토하여 제거함
        - R : MASS → stepAIC / python : dmba → stepwise_selection, forward_selection, backward_elimination
    - 벌점 회귀 : 능형(릿지)회귀, 라소회귀
    - 과적합의 문제를 해결하기 위해 교차검증과 함께 사용

4. 가중회귀
    - 복잡한 설문 분석에 중요함
    - 유용성
        - 서로 다른 관측치를 다른 정밀도로 측정했을 때 역분산 가중치를 얻을 수 있음. 분산이 높을수로 가중치가 낮음
        - 가중치 변수가 집계된 데이터의 각 행이 나타내는 원본 관측치의 수를 인코딩하도록 집계된 형식의 행이 데이터를 분석할 수 있음
    - 참고 블로그
        - 오차의 등분산성의 가정이 의심스러울 때 / 이상치의 영향을 덜 받는 회귀 모형을 만들고 싶을 때 사용할 수 있음

▼ 4.3 회귀를 이용한 예측

1. 외삽의 위험
    - 데이터 범위를 초과하면서까지 외삽하는데 사용해서는 안 됨 (시계열 분석 회귀 제외)
    - 회귀모형은 충분한 데이터 값이 있는 예측변수에 대해서만 유효함
2. 신뢰구간과 예측구간
    - 모두 변동성(불확실성)에 관한 통계값
    - 신뢰구간 : 회귀계수의 신뢰구간
        - 회귀계수 주변의 불확실성을 정량화한다.

    - 부트스트랩 알고리즘에 따라 1000개의 표본으로 회귀모델을 만들고 회귀 계수를 구한다.
    - 각각에 대해 적합한 백분위 수를 구한다.
- 예측구간 : 예측값의 신뢰구간
    - 개별 예측값의 불확실성을 정량화한다.
    - 잔차로 모델링

  관련 영상 신뢰구간 vs 예측구간

  - 신뢰구간은 fixed target / 예측구간은 moving target에 대한 불화실성을 나타냄

▼ 4.4 회귀에서의 요인 변수

요인 변수 (factor variable) = 범주형 변수(categorical variable)

지표 변수(indicator variable) = 이진 변수(binary variable)

가변수 : 요인데이터를 0과 1의 이진 변수로 부호화 한 변수

1. 가변수 표현
    - 원-핫 인코딩 : 요인 데이터를 0과 1로 표현
    - p 개의 개별 수준을 갖는 요인변수는 보통 p-1개로 표현

	property type
1	Multiplex
2	Single Family
3	Townhous

	Multiplex	Single Family
1	1	0
2	0	1
3	0	0

2. 다수의 수준을 갖는 요인변수들
    - 그대로 원핫인코딩을 하면 너무 많은 컬럼이 생김
    - 적절한 기준에 따라 그룹화해서 사용할 수 있음
3. 순서가 있는 요인 변수
    - 일반적으로 숫자값으로 변환하여 그대로 사용
    - A등급, B등급, C등급 → 3, 2, 1

▼ 4.5 회귀방정식의 해석

1. 예측변수 간 상관

- `침실의 개수`의 회귀계수가 음수가 나옴 → `집의 크기` - `침실의 개수`의 상관 관계 때문

  (예전에 제가 섹션 2에서 상관관계는 양수인데 왜 회귀계수는 음수가 나오냐?는 질문을 한 적이 있는데 이거 때문이었네요!)

- 상관관계가 있는 예측 변수를 사용하게 되면 모델 해석이 복잡해짐

2. 다중공선성
   - 변수의 상관이 극단적인 경우 다중공선성 문제가 발생
   - 다중공선성이 발생하는 경우
     - 오류로 인해 한 변수가 여러 번 포함되는 경우
     - 요인변수로부터 P-1개가 아닌 P개의 가변수가 만들어진 경우(원핫인코딩)
     - 두 변수가 서로 거의 완벽하게 상관성이 있는 경우
   - 회귀분석에서 다중공선성 문제는 반드시 해결해야하고 이 문제가 없어질 때까지 변수를 제거해야 함
     - 다중공선성 문제 해결 예제(파이썬)
3. 교란변수
   - 회귀방정식에 중요한 변수가 포함되지 못해서 생기는 누락의 문제

     → 이 상태에서 방정식 계수의 순진한 해석은 잘못된 결론으로 이어질 수 있음!
4. 상호작용과 주효과
   - 주효과 : 예측변수(독립변수)
   - 상호작용 : 두 예측변수의 상호작용 효과($x_1 * x_2$)

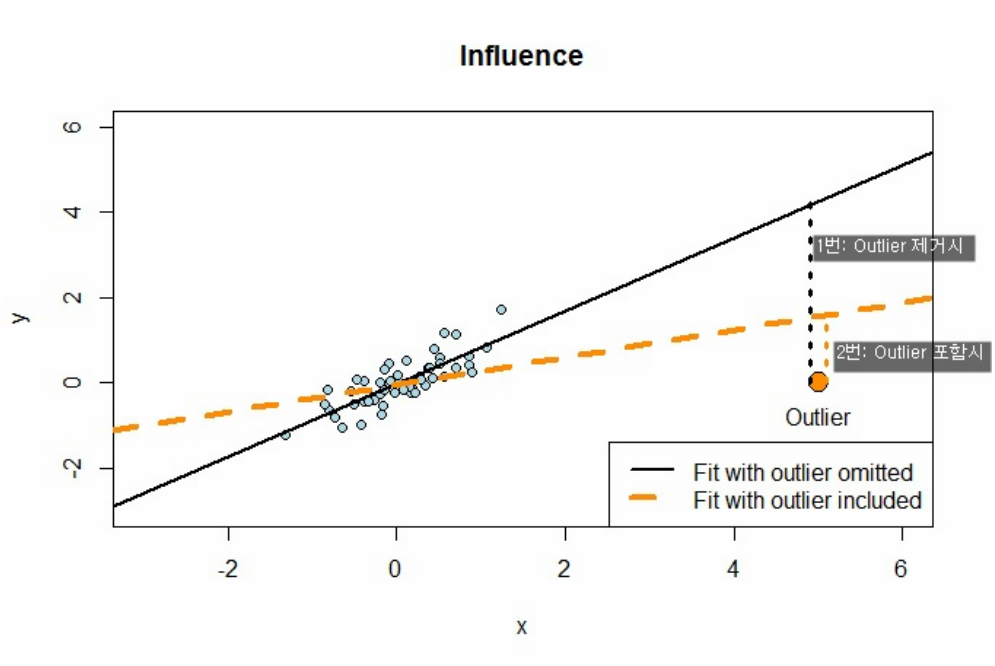
▼ 4.6 회귀진단

1. 특잇값
   - 특잇값 : (일반적) 측정치에서 멀리 떨어진 값; (회귀) 실제 y값이 예측된 값에서 멀리 떨어진 경우
   - 표준화잔차
     - 잔차를 표준오차로 나눈 값
     - 회귀선으로부터 떨어진 정도를 표준오차 개수로 표현한 값
     - 회귀에서 특잇값 발견에 사용됨
     - 빅데이터의 경우 특잇값이 큰 문제가 되지 않지만 특잇값을 찾는 것이 주목적(사기 사건 등)인 경우 이 값들이 매우 중요해짐
2. 영향값
   - 영향값 : 있을 때와 없을 때 회귀방정식이 큰 차이를 보이는 값 혹은 레코드

- 레버리지 : 회귀식에 한 레코드가 미치는 영향력의 정도(hat value)

  참고자료 : hat value 설명

  - 0 ~ 1 사이의 값을 가지며 특정 값이 영향이 있는지 없는지 확인(0: 영향 없음, 1: 영향 있음)
  - 이상값의 경우 1번 잔차값이 2번 잔차값 대비 커기게 되므로 hat value는 1에 가까워짐



$$hatvalue = 1 - \frac{outlier\text{포함잔차(아래2번)}}{outlier\text{제거잔차(아래1번)}}$$

- 쿡의 거리 : 레버리지와 잔차의 크기를 합쳐서 영향력으 판단
- 데이터 크기가 작을 경우에만 영향력이 큰 관측 데이터를 확인하는 작업이 유용함

3. 이분산성, 비정규성, 오차 간 상관 (꽤나 중요함 from 주윤님)

- 잔차(오차)의 가정

  (1) 동일한 분산을 가지며 - levene test

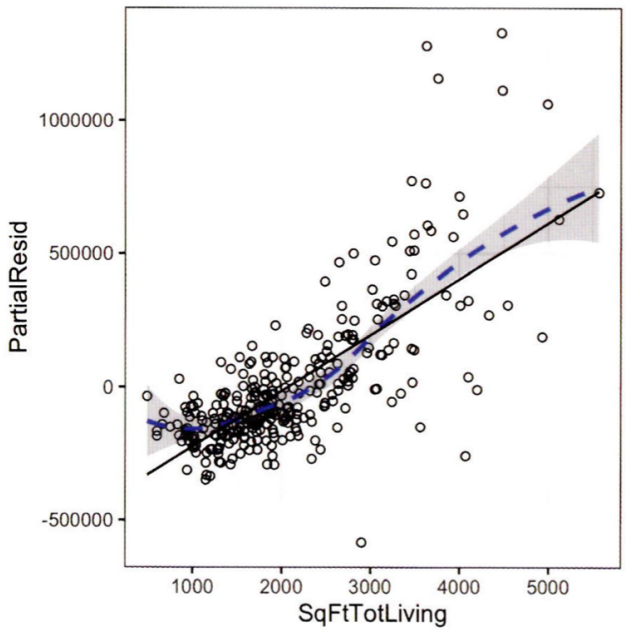
  https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html

  (2) 정규분포를 따르고 - shapiro test
  https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html

  (3) 서로 독립적이다

- 이분산성 : (1)번 가정에 위반

        - 잔차의 분산이 일정하지 않은 것
        - 모형의 불완정성이 보여줌(어떤 경우에는 맞고 어떤 경우에는 틀리다)
    - 비정규성 : (2)번 가정에 위반
        - 잔차의 분포가 정규분포와 다르고 왜곡이 있음
    - 오차 간 상관 : (3)번 가정에 위반
        - 시계열 회귀분석에서는 유의미하는 자기상관이 있는지 탐지(더빗-왓슨 통계량 사용)
        - 오차 간 상관이 있는 경우 단기 예측에 유용할 수 있으며 모델을 만들 때 함께 고려해야 함
    - 데이터 과학에서 공식 통계적 추론을 입증하기 위해 위의 가정을 만족시키는 것을 너무 신경 쓸 필요는 없다.
4. 편자차그림과 비선형성
    - 예측모델이 예측변수와 결과 변수 간의 관계를 얼마나 잘 설명하는지 시각화 하는 방법
    - 단일 예측변수를 기반으로 한 예측값 + 전체를 고려한 회귀식의 실제 잔차
    - 예시 sqfttotliving 변수가 주택 가격에 미치는 영향
        - 실제 두 변수의 관계는 비선형 → 비선형 항을 고려할 것을 생각해볼 수 있음

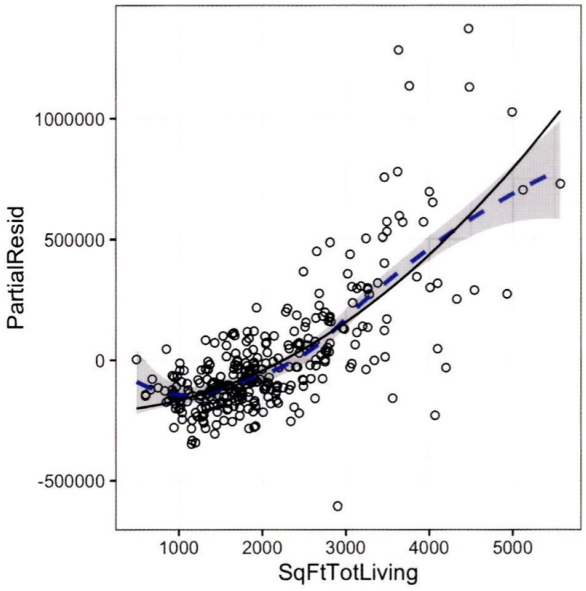


▼ 4.7 다항회귀와 스플라인 회귀

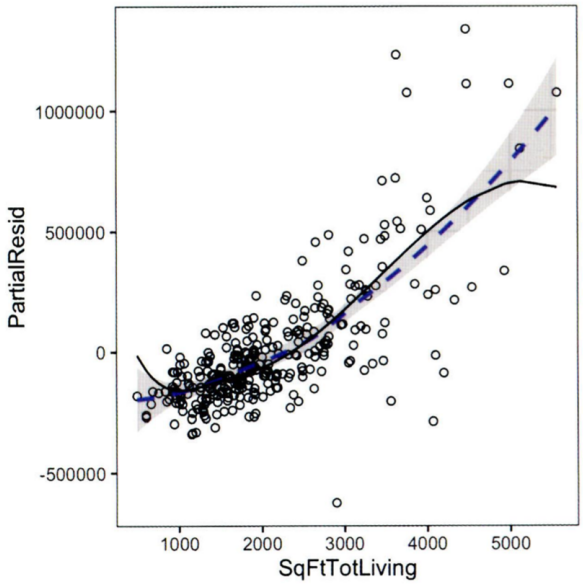
1. 다항식

- 회귀식에 다항 항을 포함한 것

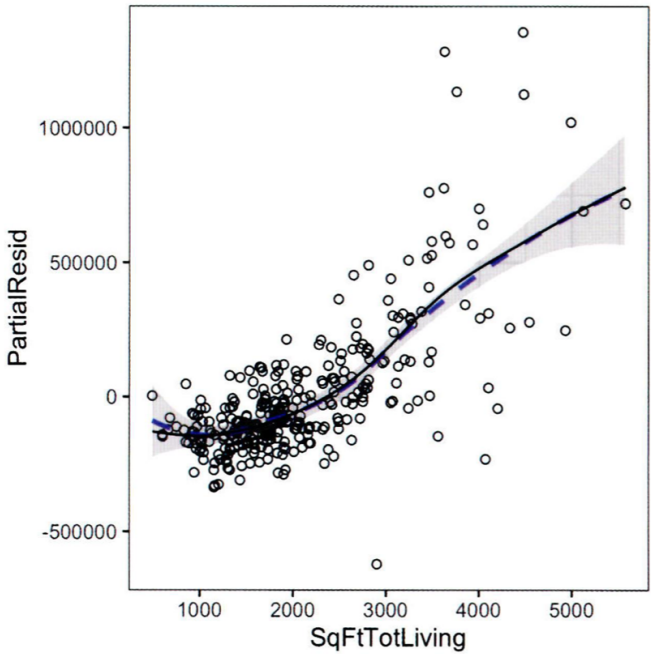
  ex)$Y = b_0 + b_1 X + b_2 X^2 + e$



2. 스플라인
   - 고정된 점들 사이를 부드럽게 보간하는 방법



3. 일반화가법모형
   - 스플라인 회귀를 자동으로 찾는데 사용할 수 있는 유동적인 모델링 기술


PartialResid
1000000
500000
0
-500000
1000
2000
3000
4000
5000
SqFtTotLiving