chapter4) 통계모델 기본

4.1 통계모델

- 모델: 모형
- 모델링: 모델을 만드는 것(통계모형을 만드는 것: 통계모델링)
  모델은 현실 세계의 이해와 예측에 활용 가능
  복잡한 세계를 단순화 할 수 있음
  분석의 목적에 맞춰서 작성하는 모델과 주목하는 관점을 바꾸는 것 가능

- **수리모델**: 현상을 수식으로 표현한 모델
  - 가상 모델 "맥주 매상은 기온에 의해 변한다 "
  - $$\text{맥주매상}(\text{만원}) = 20 + 4 * \text{기온}(\text{섭씨})$$
    '수식으로 표현 시,맥주와 기온의 관계를 더욱 명확하게 표현 가능'

- **확률모델**: 수리모델 중에서도 확률적인 표현이 있는 모델
  일반적으로는 정규분포 사용(2가지 방법 존재)
  - 정규분포 가정 시, 맥주 매상을 기온으로 설명하는 확률모델
    방법 1.
    $$\text{맥주매상} \sim N(20 + 4 * \text{기온}, \sigma^2)$$
    *'맥주 매상은 평균이 20+4 * 기온이고, 분산이 \sigma^2인 정규분포를 따른다'*

    방법 2.
  - $$\text{맥주매상} = 20 + 4 * \text{기온} + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$
    *'맥주 매상은 20+4 * 기온에 대해 평균이 0, 분산이 \sigma^2인 정규분포를 따르며 노이즈가 있다'*

- **통계모델**: 데이터에 적합하게 구축된 확률모델
  $$\text{맥주매상} \sim N(10 + 5 * \text{기온}, \sigma^2)$$
  확률모델의 구조를 생각하면서 데이터에 적합하게 파라미터를 조정해가며 통계모델 구축
  확률모델과 통계모델의 구별은 엄밀하지 않아 같은 의미로 사용되는 경우도 있음

- 확률분포와 통계모델
  통계모델을 사용하면 확률분포의 모수(파라미터) 의 변화 패턴을 명확히 할 수 있음

- 통계모델을 이용한 예측
  다음과 같이 통계모델 추정,

$$\text{맥주매상} \sim N(10 + 5 * \text{기온}, \sigma^2)$$

= "기온이 10도일 때의 매상 예측은 '기댓값이 60, 분산이 \sigma^2인 정규분포를 따르는 매상 데이터를 얻을 것' 이라는 주장

- 통계모델에 의한 예측은 기온이라는 독립변수를 얻는 것이 조건인 매상의 확률분포, 즉 조건부 확률분포의 형태로 얻을 수 있음
- 예측값의 대푯값을 1개 고르는 경우, 조건부 기댓값이 사용됨(위의 경우, 대푯값은 60)

- 통계모델과 고전적인 분석 절차의 비교
  - 고전적인 평균값의 차이 검정 등은 통계모델의 활용 방법 중 한 가지에 불과함
    - ex) 상품의 가격과 매상의 관계 조사(전히 똑같은 상품에 대해 가격이 쌀 때와 비쌀 때의 매상 평균값을 비교해서 매상에 유의미한 차이가 있는지 검정) -> **평균값 차이 검정으로 대응**

      모델1: 가격이 쌀 때와 비쌀 때의 매상 평균값은 변하지 않는다.

      모델2: 가격이 쌀 때와 비쌀 때의 매상 평균값은 변한다.

      평균값의 차이 검정은, '1단계: 2가지 모델을 작성'과 '2단계: 어느 모델이 더 들어맞는지 판단하기'라는 두가지 작업

      그 중, 2단계의 판단만 사람들에게 보이게 됨.

- 통계모델의 활용

  모델을 구축하고 결과를 통해 다양한 결과 얻을 수 있음

  그러나 추정된 모델 안에서만 성립하는 결과라는 점 주의해야 함(통계모델은 '잠정적인'세계의 모형일 뿐)

  가령, 통계모델 구축 시, 파라미터의 추정을 완전히 틀리게 하면 올바른 해석 할 수 없음

  그러나, 통계모델은 현대 데이터 분석의 표준 도구라고 할 수 있음

4.2 통계모델을 만드는 방법

[예제: 맥주 매상 예측 모델]

맥주 매상에 영향 주는 것(3가지): 기온, 날씨(맑음, 흐림, 비), 맥주 가격

- 종속변수와 독립변수
  - 종속변수(응답변수): 어떤 요인에 의해 종속된 변수, 어떤 변화에 응답하는 변수 -> 맥주 매상
  - 독립변수(설명변수): 흥미 있는 대상의 변화를 설명하는 변수, 모델 내의 다른 대상에 영향 받지 않는(독립적인) 변수 -> 기온, 날씨, 맥주가격
    - 독립변수를 사용해서 종속변수를 모델링 함 (방향성 존재!)
    - 확률모델에서는 '종속변수~독립변수'로 표기하는 경우가 많음.
- 파라메트릭 모델

  :가능한 한 현상을 단순화해서 소수의 파라미터만 사용하는 모델
- 논파라메트릭 모델

  : 소수의 파라미터만 사용한다는 방침을 취하지 않는 모델 (이 책에서는 사용X)

- 선형 모델

  : 종속변수와 독립변수의 관계를 선형으로 보는 모델

  - 맥주 매상과 기온의 관계를 선형이라고 가정하여 모델링

    $$\text{맥주매상}(\text{만원}) = 20 + 4 * \text{기온}(\text{섭씨})$$

    위 모델은 기온이 1도 오르면, 매상이 4만원 오른다고 생각함 -> 이러한 변하지 않는 관계를 가정한 것이 선형(P.226참고)
- 계수와 가중치
  - 계수: 통계모델에 사용되는 파라미터
  - 기온만 사용해서 맥주 매상을 예측하는 모델

    $$\text{맥주매상} \sim N(\beta_0 + \beta_1 * \text{기온}, \sigma^2)$$

    $$\beta_n : \text{계수} or \beta_0 : \text{절편} \beta_1 : \text{회귀계수}$$

    계수와 독립변수(여기서는 기온)가 있으면 모수(여기서는 정규분포의 평균값)를 추측할 수 있음

    통계학에서는 계수라고 부르지만, 머신러닝에서는 같은 내용을 나타내는 데 가중치라는 표현을 사용하는 경우도 있음

- **모델 구축 = 모델 정하기 + 파라미터 추정**
- 모델 구축 작업
  1. 모델의 특정(모델의 구조를 수식으로 표현하는 것)

     모델의 구조란, '기온이 변화하면 맥주 매상이 증가하거나 감소한다'와 같은 구조
  2. 파라미터(계수) 추정

     '기온이 1도 오르면 맥주 매상이 x만원 증가한다'에서 x만원 부분을 추정하는 것

  따라서, 모델을 구축할 때는 '모델의 구조'와 '파라미터' 두 가지를 결정해야 함.
- 선형모델을 구축하는 방법

  모델 구축의 두 가지 작업 중 파라미터 추정은 파이썬 이용해서 대부분 자동으로 완료
  - 선형모델임을 가정했을 때, 모델의 구조를 바꾸는 방법
    1. 모델에 사용되는 독립변수를 바꾼다
    2. 데이터가 따르는 확률분포를 바꾼다
- 변수 선택

  : 모델에 사용될 독립변수를 고르는 작업

  변수 선택을 위해서는 우선 여러 가지 변수 조합 모델을 만들어봐야 함

  ex) 독립변수 A, B, C가 있는 경우
  - 종속변수 ~ 독립변수 없음
  - 종솝변수 ~ A
  - 종속변수 ~ B
  - 종속변수 ~ C
  - 종속변수 ~ A+B
  - 종속변수 ~ A+C
  - 종속변수 ~ B+C
  - 종속변수 ~ A+B+C

종속변수: 맥주 매상, 독립변수: 기온, 날씨, 맥주 가격

(독립변수가 없을 경우의 모델은 맥주 매상의 평균이 언제나 일정하다고 가정한 모델이라고 해석가능)

위의 변수 조합들 중 가장 좋은 변수의 조합을 가진 모델을 선택하는 것이 변수 선택.

- 가장 좋은 변수 조합 선택 방법
    1. 통계적 가설 검정 이용
    2. 정보 기준 이용

- Null 모델

  : 독립변수가 없는 모델 (null:아무것도 없다는 뜻)

- 검정을 이용한 변수 선택

$$\text{맥주매상} \sim N(\beta_0 + \beta_1 * \text{기온}, \sigma^2)$$

통계적 가설 검정을 이용하는 경우, 아래와 같은 가설 세움

- $\text{귀무가설 : 독립변수의계수}\beta_1\text{은}0\text{이다.}$
- $\text{대립가설 : 독립변수의계수}\beta_1\text{은}0\text{이아니다.}$

  귀무가설이 기각되는 경우 -> 기온에 대한 계수가 0이 아니라고 판단 가능하므로, 모델에 기온이라는 독립변수가 필요함

  귀무가설을 기각할 수 없을 때는 모델은 간단한 쪽이 좋다는 원칙에 의해 변수를 모델에서 제거함. 이 경우, 유일한 독립변수가 제거되기 때문에 Null모델이 됨.

  (분산분석이라는 검정 방법은 5장에 파이썬 을 이용한 구현을 할 때 알아볼 것)

- 정보 기준을 이용한 변수 선택

  모델 선택의 또 다른 방법은 정보 기준(추정한 모델의 좋은 정도를 정량화한 지표, 아케이케 정보 기준 등이 자주 사용)을 사용하는 것

  아케이케 정보 기준(AIC)가 작을 수록 좋은 모델이라고 판단.

  모델에서 가능한 변수의 패턴을 망라하여 모델을 구축하고, 각 모델의 AIC를 비교.

  AIC가 가장 작은 모델을 채택함으로써 변수 선택을 실행함

- 모델 평가
- : 추정된 모델을 평가하는 단계
    - 평가 관점
        1. 예측 정확도의 평가
        2. 모델을 구축할 때 가정한 전제조건을 만족했는지 체크

            EX) 맥주 매상을 아래와 같이 모델링

$$\text{맥주매상} = 20 + 4 * \text{기온} + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

이때 모델의 전제조건이 만족되었다면 매상의 예측값과 실젯값의 차이 \varesilon은 평균이 0인 정규분포를 따르게 될 것.

이러한 부분을 체크하는 것이 모델 평가. (BUT, 5장에서 파이썬으로 구현하며 설명할 것임)

- 통계모델을 만들기 전 분석 목적 정하기

  실제로 파이썬 코드를 작성하기 전에 분석의 목적을 결정하고 데이터를 수집하여 모델링하는 것이 중요함

  ex) 매상을 늘리는 것이 목적이라며 기온과 맥주의 관계를 모델링하는 것은 우리가 기온을 변화시기 어렵기 때문에 무의미함

  그러나, 기온을 바탕으로 매상을 예측하여 재고관리에 활용하려는 것이 목적이라면 이 모델은 도움이 됨

4.3 데이터의 표현과 모델의 명칭

- **정규선형모델**

  : 종속변수가 정규분포를 따르는 것을 가정한 선형모델(파라메트릭모델)

  $$\text{종속변수범위} : -\infty \sim +\infty, (\text{종속변수는연속형변수})$$

- **회귀분석(회귀모델)**

  : 정규선형모델 중 독립변수가 연속형 변수인 모델

- **다중회귀분석**

  : 독립변수가 여러개 있는 것(이와 대비되는 독립변수가 1개인 회귀분석은 단일회귀분석)

- **분산분석**

  : 정규선형모델 중에서 독립변수가 카테고리형 변수인 모델

  (이 책에서 분산분석은 항상 검정 방법을 가리키는 이름으로 사용)

  > 일원분산분석: 독립변수가 1종류일 때
  >
  > 이원분산분석: 독립변수가 2종류일 때

- **일반선형모델**

  :종속변수가 따르는 확률분포를 정규분포 이외의 분포에도 사용 가능한 선형모델 (6장)

- **머신러닝에서의 명칭**

  머신러닝 분야에서 회귀는 종속변수가 연속형 변수인 모델 의미함. 이 경우 정규선형모델은 넓은 의미에서 회귀가 됨.

  종속변수가 카테고리형 변수인 모델은 분류 또는 식별모델이라고 함.

  ex) 모집단 분포를 이항분포라고 가정 시, 식별모델이며 포아송분포라고 가정 시, 회귀모델이 됨(분야에 따라서 용어 다르므로 주의)

4.4 파라미터 추정: 우도의 최대화

파라미터 추정은 파이썬 함수를 사용하면 간단함. *따라서 계산 방법보다는 계산의 의미나 해석에 중점을 두었음*

- 파라미터 추정 방법 배우는 의미

  파라미터 추정 원리 몰라도 파이썬 사용하면 통계모델 구축하고, 예측이나 현상에 이용 가능함

  구조이해라는 면에서 의미 있으나, 세세한 알고리즘까지 이해할 필요 없음.

- 우도

  : 파라미터가 정해져 있을 때 표본을 얻을 수 있는 확률(밀도)을 우도라고 함. (L로 표기)

  EX) 동전 앞 나올 확률 1/2 -> 이때, 1/2가 파라미터임

  동전 2번 던져서 첫번째는 앞면 두번째는 뒷면 나옴 -> 이것이 표본

  이 표본을 얻을 수 있는 확률 1/2*1/2=1/4

  앞면이 나올 확률이 1/3인 동전의 경우, 우도는 1/3*2/3=2/9.

- 우도함수

  : 파라미터를넘겨서 우도를 계산할 수 있는 함수

  - 동전을 던져서 앞면이 나올 확률을 파라미터로 가정.

    $$\theta\text{를 파 라 미 터 로 가 정 }, \theta\text{를 지 정 하 여 우 도 를 구 하 는 우 도 함 수 } = L(\theta)$$

    이때 우도함수는

    $$L(\theta) = \theta * (1 - \theta)$$

- 로그우도

  :로그에 우도 취한 것

- 로그의 성질(P.234~236)

- 최대우도법

  : 우도나 로그우도의 결과를 최대로 하기 위한 파라미터를 추정할 때 사용하는 방법

  - 동전던지기 예시

    $$\text{파 라 미 터 } \theta\text{가 } 1/2\text{일 때 우 도 는 } 1/4\text{입 니 다 .}$$

    $$\text{파 라 미 터 } \theta\text{가 } 1/3\text{일 때 우 도 는 } 2/9\text{입 니 다 .}$$

    $$1/4\text{과 } 2/9\text{중 에 서 } 1/4\text{가 크 기 때 문 에 } \theta\text{는 } 1/2\text{이 좋 다 고 할 수 있 습 니 다 .}$$

- 최대우도추정량

  : 최대우도법에 의해 추정된 파라미터

  추정값임을 나타내기 위해 햇(모자)기호를 사용하여,

  $$\hat{\theta}\text{로 표 시}$$

- 최대화 로그우도

  : 최대우도추정량을 사용했을때의 로그우도

  $$logL(\hat{\theta})$$

- 정규분포를 따르는 데이터의 우도

  모집단이 정규분포를 따른다고 가정했을 때의 최대우도법 계산 예

- 독립변수가 없는 Null모델의 파라미터 추정 방법

  맥주 매상 : 변수y

  $$y \sim N(\mu, var^2), 평균\ \mu, 분산\ \sigma^2 인\ 정규분포따름$$

  샘플사이즈가 클수록 좋지만, 계산을 간단하게 하기 위해 샘플사이즈가 2인 표본으로 계산 방식을 설명

  $$y_1을\ 얻었을\ 때의\ 확률밀도\ = N(y_1|\mu, var^2)$$

  $$y_2을\ 얻었을\ 때의\ 확률밀도\ = N(y_2|\mu, var^2)$$

  이때의 우도는, 이를 최대로 하는 파라미터 \mu와 \var^2을 계산하면 됨

  $$L = N(y_1|\mu, var^2) * (y_2|\mu, var^2)$$

- 장애모수

  : 직접적인 관심의 대상이 아닌 파라미터

  정규분포의 모수는 평균과 분산 2가지. 그러나 분산은 평균값에서 계산할 수 있으므로 평균을 추정할 수 있으면 분산 또한 덩달아서 알 수 있게 됨. 그러므로 분산이라는 파라미터에는 관심을 두지 않고 이미 알고 있는 것으로 취급

  정규분포를 추정할때도 최대우도법에서는 분산을 장애모수로 취급함. 따라서 Null모델의 경우 평균만 추정하면 됨

- 정규선형모델의 우도

  최대우도법에 의한 파라미터 추정을 맥주 매상 모델과 연결지어 설명

  (종속변수는 정규분포를 따른다고 가정하므로 정규선형모델로 간주 가능)

  $$맥주\ 매상\ \sim N(\beta_0 + \beta_1 * 기온, \sigma^2)$$

  - 계수 beta를 결정했다고 하고, 그 때의 우도를 계산

    샘플사이즈 2인 표본이라고 가정, 맥주 매상은 y, 그날의 기온은 x로 표기.

    우도는 아래와 같음(이때, 분산은 장애모수)

    $$L = N(y_1|\beta_0 + \beta_1 x_1, \sigma^2) * N(y_2|\beta_0 + \beta_1 x_2, \sigma^2)$$

    좀 더 일반적인 샘플사이즈 N인 표본에 대한 우도

    $$L = \prod_{i=1}^{N} N(y_i|\beta_0 + \beta_1 x_i, \sigma^2)$$

    여기에 로그를 취하면,

    $$logL = \sum_{i=1}^{N} log[(N(y_i|\beta_0 + \beta_1 x_i, \sigma^2)]$$

    로그우도를 최대로 하는 파라미터 \beta_0과 \beta_1을 추정치로 사용하는 것이 최대우도법

    함수의 결괏값을 최대로 하는 파라미터를 구하는 것을 arg max라고 쓰며, 최종적으로는 아래와 같이 정리

    $$\underset{\beta_0,\beta_1}{arg\,m}axlogL = \underset{\beta_0,\beta_1}{arg\,m}axlogL[(N(y_i|\beta_0 + \beta_1 x_i, \sigma^2)]$$

    - N() 대신 정규분포의 확률밀도 함수를 넣으면 로그우도를 계산 가능

      계산식이 어렵다고 느껴지면 중간 과정 무시하고 결과만 보아도 상관없음.

$$\arg\max_{\beta_0,\beta_1} \log \mathcal{L}$$

$$= \arg\max_{\beta_0,\beta_1} \sum_{i=1}^{N} \left[\log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \right\} \right] \right]$$

$$= \arg\max_{\beta_0,\beta_1} \sum_{i=1}^{N} \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left[\exp\left\{ -\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \right\} \right] \right]$$

$$= \arg\max_{\beta_0,\beta_1} \sum_{i=1}^{N} \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \right]$$

- 최대우도법 계산 예

  최대우도법은 미분 사용해서 해석적으로 해를 얻을 수 있음

  그러나 이 계산은 이 책 다른 곳에서는 다루지 않으므로 어렵다면 넘어가도 괜찮음.

  - Null 모델을 대상으로 계산

    파라미터 \mu추정

$$\text{맥주매상} \sim N(\mu, \sigma^2)$$

$$arg\underset{\beta_0,\beta_1}{max} logL$$

$$= arg\underset{\beta_0,\beta_1}{max} \sum_{i=1}^{N} log[(N(y_i|\mu, \sigma^2)]$$

$$= arg\underset{\beta_0,\beta_1}{max} \sum_{i=1}^{n} [log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(y_i - \mu)^2}{2\sigma^2}]$$

    최댓값을 구하는 경우에는 미분해서 값이 0이 되는 점을 찾는 것이 정석임.

    이번에는 \mu를 변화시켜서 로그우드가 최대가 되는 점 찾기(\mu에서 로그우도함수를 미분했을 때 그 값이 0이 되는 \mu를 찾으면 됨)

$$\sum_{i=1}^{n} [\frac{2(y_i - \mu)^2}{2\sigma^2}] = 0$$

    분산은 장애모수이므로 상수 취급하여 지울 수 있음

$$\sum_{i=1}^{n} [(y_i - \mu] = 0$$

    \mu를 시그마 밖으로 빼냄

$$\sum_{i=1}^{n} [y_i] - N\mu = 0$$

$$\mu = \frac{1}{N} \sum_{i=1}^{n} y_i$$

    => Null 모델의 로그우도를 최대로 하는 파라미터 \mu는 종속변수의 평균값과 같다는 뜻이 됨!

- 최대우도추정량의 성질

  최대우도추정량은 추정오차라는 측면에서 매우 바람직함

  -> 샘플사이즈가 한없이 커질 때 추정량의 표본분포가 점근적으로 정규분포를 따름(점근적 정규성)

점근적 정규성을 갖는 추정량 중 점근 분산을 최소로 하는 추정량으로도 알려져 있으므로 최대우도법은 점근 유효추정량임.

표본분산의 분산이 작다는 것은 추정치의 흩어짐이 작고, 추정의 오차가 작다는 의미이므로 최대우도추정량은 바람직한 성질을 가진 추정량임.

4.5 파라미터 추정: 손실의 최소화

파라미터 추정의 기본 개념은 모델에 잘 들어맞는 파라미터를 채용한다는 것

최대우도법은 모델에 들어맞는 정도를 우도로 수치화해서 그것이 최대가 되는 파라미터를 추정했음

이 장에서는 머신러닝에서 자주 사용되는 개념인 손실의 최소화라는 측면에서 파라미터를 추정하는 방법 살펴봄

- ***손실함수**

  : 파라미터 추정을 할 때 손실을 최소화하는 목적으로 사용

- **잔차**

  : 실제 종속변수 값과 모델을 이용해 계산한 종속변수의 추정치와의 차이

  $$\text{맥주매상} \sim N(\beta_0 + \beta_1 * \text{기온}, \sigma^2)$$

  예를 들어 기온이 20도 였을 때 맥주 매상의 기댓값은 아래와 같이 계산

  $$\beta_0 + \beta_1 * 20$$

  이것이 기온 20도일 때 맥주 매상의 추정치(점추정치)임.

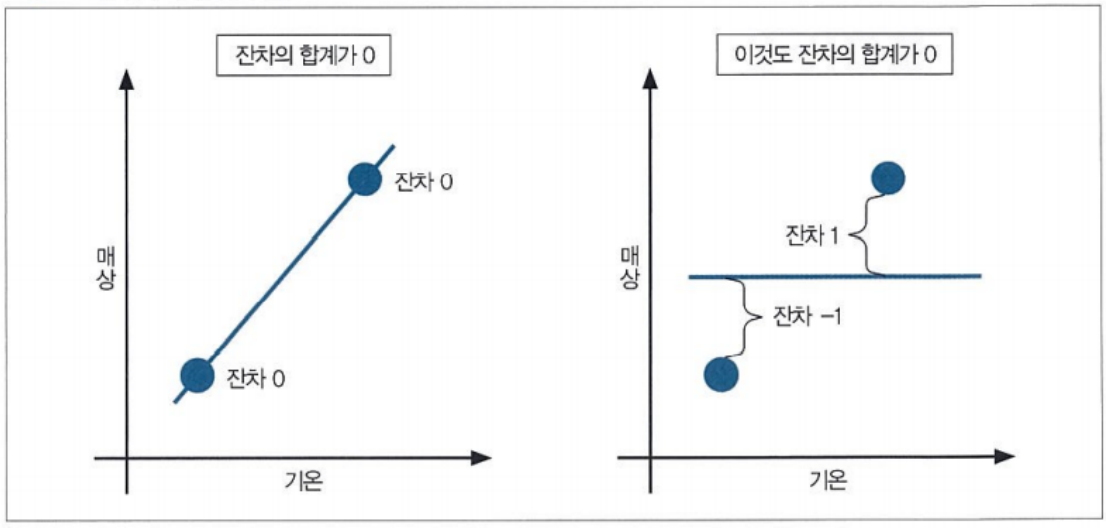
  종속변수(여기서는 맥주 매상)을 y라고 하고, 모델에 의한 종속변수의 추정치는 \widehat y이라고 하면, 잔차는 아래와 같음

  $$residuals = y - \hat{y}$$

- 잔차의 합을 그대로 손실의 지표로 사용할 수 없는 이유

  -> 잔차의 합이 0이 되기 때문 (p.242)

그림 4-2 잔차가 가지는 문제점



- **잔차제곱합**

  : 잔차를 제곱해서 합계를 구한 것(잔차 합이 가지는 문제 해결)

$$\text{잔차제곱합} = \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2]$$

- **최소제곱법**

  : 잔차제곱합을 최소로 하는 파라미터를 채용하는 방법

  손실함수로 잔차제곱합을 사용하여 손실을 최소화하는 파라미터를 추정치로 하는 방법이라고 얘기할 수 있음(Ordinary Least Squared, OLS)

- **최소제곱법과 최대우도법의 관계**

  최소제곱법 이용한 파라미터 추정치는 모집단 분포 정규분포 가정 시, 최대우도법 결과와 일치함 (증명 P.244)

- **오차함수**

  :머신러닝 분야에서 로그우도의 부호를 바꾼 것

  로그우도의 플러스마이너스를 바꾼 것이므로 이를 최소화하는 것은 우도를 최대로 하는 것과 같음

- **여러가지 손실 함수**

  잔차제곱법으로 최대 우도법을 이용한 추정치와 같은 파라미터 추정 가능

  그러나 모집단 분포가 정규분포 이외의 분포임을 가정할 경우, 최대우도법의 추정치와 최소제곱법의 추정치는 일치하지 않음. 따라서 데이터에 따라서 손실함수를 바꾸어야 함

4.6 예측 정확도의 평가와 변수 선택

- 적합도와 예측 정확도
  - 적합도 : 가지고 있는 데이터에 대해 모델을 적용했을 때 들어맞는 정도
  - 예측 정확도: 아직 얻지 못한 데이터에 대해 모델을 적용했을 때 들어맞는 정도
- 과적합(오버피팅)

  : 적합도는 높은데, 예측 정확도가 낮아지는 경우

- 변수 선택의 의의

  :과적합을 하게 되는 흔한 원인으로는 독립변수를 너무 많인 늘리는 경우

  즉, 필요 없는 독립변수를 제외하는 것만으로도 예측 정확도가 높아질 가능성 존재

- 일반화 오차

  : 아직 얻지 못한 데이터에 대한 예측오차

- 훈련 데이터와 테스트 데이터
  - 훈련 데이터(트레이닝 데이터): 파라미터 추정에 사용되는 데이터, 훈련 데이터의 정확도를 평가하는 것으로 모델의 정확도(정밀도)는 구할 수 있지만, 일반화 오차를 평가하는 건 어려움
  - 테스트 데이터: 일반화 오차를 평가하기 위해 파라미터 추정을 할 때 사용하지 않고 남겨둔 데이터, 파라미터 추정 시 사용하지 않은 테스트 데이터로 모델의 정확도를 평가하는 것으로 일반화 오차를 어느 정도 평가 가능

- **교차검증**
  - 교차검증(크로스 밸리데이션): 데이터를 일정한 규칙에 따라 훈련 데이터와 테스트 데이터로 나누어 테스트 데이터에 대한 예측 정확도를 평가하는 방법

- 리브-P-아웃 교차검증

  가지고 있는 데이터에서 p개의 데이터를 추출하고 남은 데이터를 테스트 데이터로 사용하는 방법 -> 가지고 있는 데이터 중 2개 추출해 훈련 데이터로 이용하고 나머지 데이터로 예측 정확도 평가

- K겹 교차검증

  가지고 있는 데이터를 K개의 그룹으로 나눔. 그 그룹 중에서 하나를 추출하여 테스트 데이터로 사용함. 이것을 K번 반복하여 예측 정확도의 평균값을 평가값으로 사용함

- **아카이케 정보 기준(AIC)**

$$AIC = -2 * (\text{최대로그우도} - \text{추정된파라미터수})$$

AIC가 작을수록 좋은 모델

로그우도가 클수록 적합도가 높음. 그러나, 적합도를높게 하는데만 주력할 경우 일반화 오차 커짐. 때문에 AIC로 추정한 파라미터 수를 적합도에 대한 페널티로 사용함.

독립변수가 많아지면 로그우도가 커지지만 그와 동시에 페널티도 커짐

'AIC는 페널티를 보강하고 지나치게 로그우도가 높아지는지 판단하는 지표'

AIC를 사용하면 불필요한 변수를 제외할 수 있음. 또한 교차검증에 비해 계산량이 적다는 장점 존재

- **상대 엔트로피** (AIC 지표의 해석)
  - AIC는 통계모델의 예측을 중요시함. 진짜 분포와 통계모델에 의해 얻어진 분포의 차이를 측정하는 지표가 **상대 엔트로피**임
    - 상대 엔트로피

      : 분포 사이의 의사 거리(g(x)와 f(x)는 확률밀도 함수)

$$\text{상대엔트로피} = \int g(x) log \frac{g(x)}{f(x)} dx$$

$$
상대 엔트로피=\int g(x){{log \:g(x)- log\:f(x)}}dx

확률밀도함수에서기댓값을구하는식을다시보면,

x의 기댓값=\int f(x)*xdx
$$

상대 엔트로피는 2개의 확률밀도함수의 로그 차를 상대 엔트로피의 기댓값으로 간주하고, 이 값을 이용해 확률분포의 차이를 측정하는 지표임

- 상대 엔트로피의 최소화와 평균 로그우도

  진짜 분포와 예측 분포의 거리를 줄이는 방법

  (y는 종속변수, g(y)가 진짜 분포, f(y)가 모델로 예측한 분포)

$$\int g(y)\{log\, g(y) - log\, f(y)\}dy$$

$$\int g(y) log\, g(y) - g(y) log\, f(y) dy$$

여기서 진짜 분포 g(y)는 변경 불가. 따라서 이 거리를 줄이려면 아래 식을 최소로 하면 됨

$$\int -g(y) log f(y) dy$$

여기에 -1을 곱한 값을 **평균로그우도**라고부름

log f(y)에 있는 f(y)는 예측된 종속변수의 확률분포임.

진짜 분포와 추정된 분포의 차이를 최소화하는 것은 평균로그우도에 -1을 곱한 것을 최소화하는 것과 같은 의미 -> 평균로그우도를 최대화하는 것으로 진짜 분포와 예측된 분포의 차이를 최소화할 수 있음.

- 평균로그우도가 지니는 편향과 AIC

  평균로그우도를 구하는 것은 어려우므로 최대로그우도를 대신 사용함

  그러나 최대로그우도는 크게 편향된 값을 지니므로 문제가 생김

  따라서 이 편향을 없앤 것이 AIC이며, 아래와 같이 계산

  $$AIC = -2 * (\text{최대로그우도} - \text{추정된파라미터수})$$

  최대로그우도에서 추정된 파라미터 수를 빼는 것은 이러한 이유 때문임

- AIC와 교차검증

  모델의 적합도로 최대우도 사용, 예측의 좋음 평가에 로그우도 사용한다는 조건 만족 시, 리브-1-아웃 교차검증을 이용한 결과도 AIC최소 기준(AIC를 최소로 하는 것)에 의한 변수 선택의 결과는 점근적으로 일치하는 것으로 알려져 있음

- AIC를 이용한 변수 선택

  AIC는 모델의 좋음을 평가하는 지표(AIC가 적을수록 좋은 모델)

  -> AIC가 최소가 되는 변수의 조합을 선택하는 것

- 검정 대신 변수 선택

  EX) 약을 먹었을 때 체온이 오른다고 볼 수 있는지 알고 싶을 때

  - 모델 구축

    모델1: 체온~독립변수 없음

    모델2: 체온~ 약의 유무

    그리고 모델 1과 모델2를 각각 추정하여 AIC계산하고, 모델2의 AIC가 더 작다면 모델2를 채택하게 됨(약의 유무는 모델에 들어가는게 좋다고 판단 가능)

  그러나, 모델 2가 올바르다는 보증은 AIC가 해줄 수 없음. 가지고 있는 데이터를 이용해서 모르는 데이터에 대한 예측을 좋게 하는 것을 최대로 할 목적으로 AIC사용하는 것임

- 검정과 AIC중 어느 것을 선택할 것인가

  검정 VS AIC 중 뭐가 더 좋은지 판별 불가.

  그러나 두 가지 모두 해석을 할 수 있게 됨. *단, 지표를 바꾸지 말아야 함*