5장 재표집법

Date	@2025년 5월 8일

🔍 핵심 개념 정리

재표집법

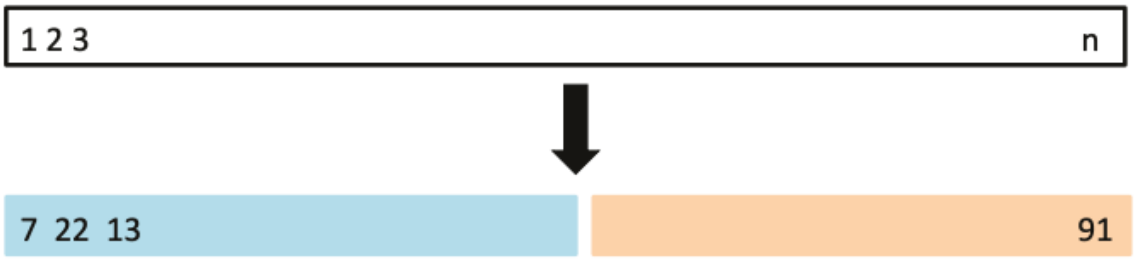
- 재표집법 : 관측된 기존 데이터를 반복적으로 샘플링하여 새로운 데이터셋처럼 사용하는 통계적 기법
- resampling 하는 이유
  - 불확실성 추정 : 모수 추정값(표준오차, 분산 등)의 정확도 추정
  - 모델의 성능 평가 : 모델의 예측 정확도 평가
- 여러 번의 샘플링으로 적합하는 과정을 거치기 때문에 계산량이 많으나 컴퓨터 성능 증가로 크게 고려 사항 아님
- 리샘플링 방법 2가지
  - 교차검증 : 모형평가와 모형선택에 사용
    - 테스트 오류를 추정해 성능 평가 → 모형평가
    - 적절한 유연성 수준 평가 → 모형선택
  - 부트스트랩
    - 모수 추정값의 정확도나 통계적 학습 방법의 정확도 측정

5.1 교차검증

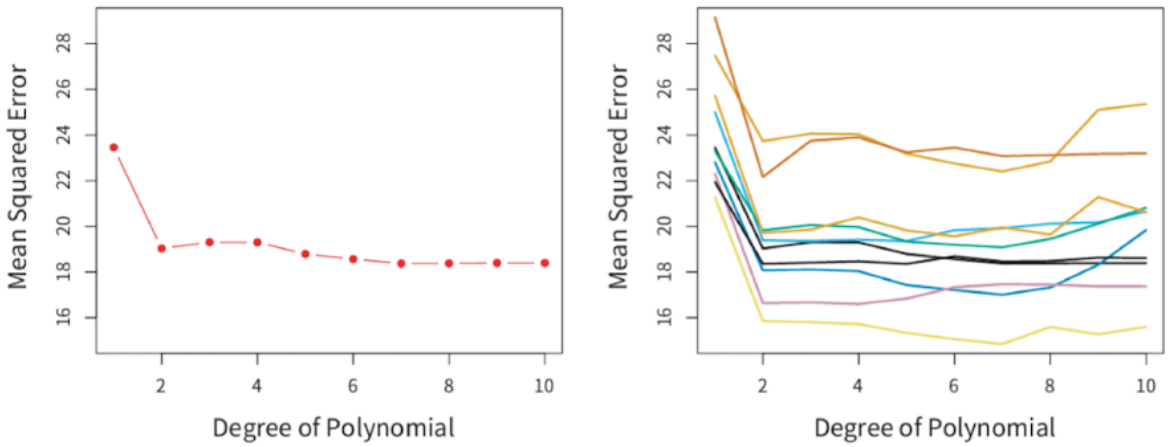
- 테스트 오류 : 새로운 관측에 대한 예측 결과의 평균 오류
- 훈련 오류 : 훈련에 사용된 관측에 대한 예측 결과의 평균 오류
- 테스트 오류율과 훈련오류율이 다르게 나타나는 경우가 많음.

- 테스트 오류를 측정하기 위해서는 테스트 셋이 확보되어야 하지만, 테스트 셋이 없는 경우 훈련 데이터를 이용해 테스트 오류율을 추정
- 훈련 관측값의 일부를 빼놓고 빼놓았던 훈련 관측값을 마치 테스트 셋처럼 사용하여 통계적 학습 방법 적용해 테스트 오류율 추정

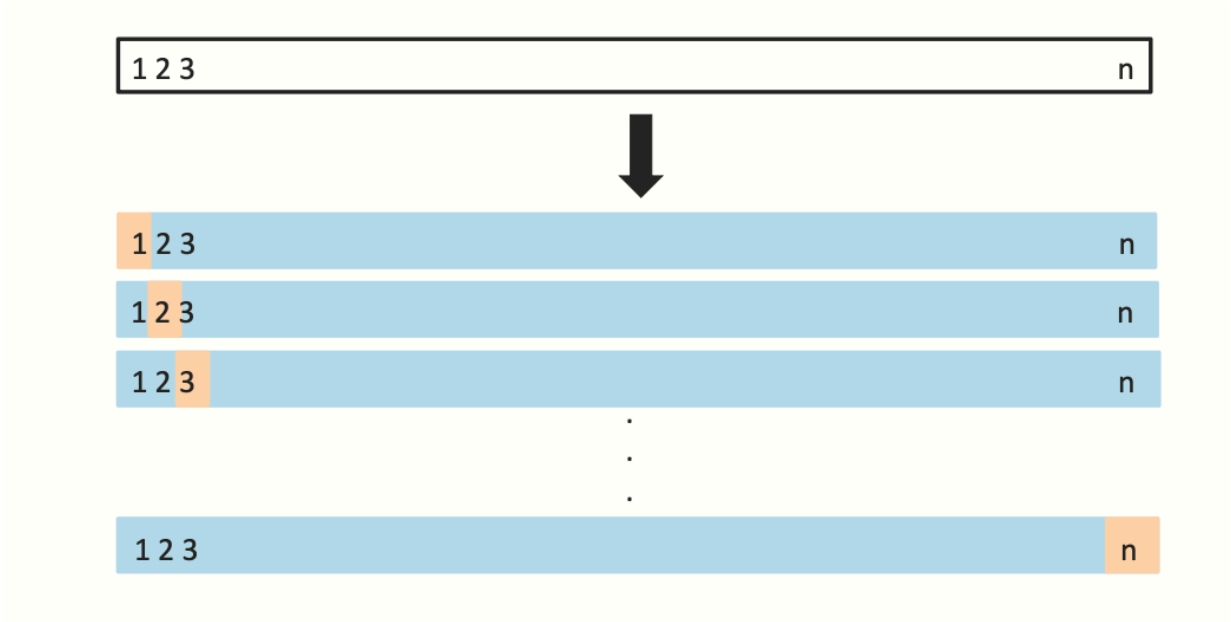
5.1.1 훈련/검증 세트 분할법



- 사용 가능한 관측값을 무작위로 훈련 세트와 검증 세트로 나눔
- 훈련 세트는 모델 학습에 사용, 검증 세트는 성능 평가에 사용
- 훈련 세트로 모형 적합 후, 해당 모형을 이용해 검증 세트로 예측하여 검증 세트의 오류율로 테스트 오류율 추정
- 장점 : 단순하고 실행이 쉬움
- 단점
  1. 테스트 오류율의 검증 추정값, 즉 검증 세트의 오류율이 검증 세트의 데이터에 따라 변동이 심함
     - 샘플링에 따라 검증 세트의 MSE 값이 크게 변화
     - 훈련 세트에서 훈련한 관측값과 유사한 데이터가 검증 세트에 포함되어 있으면 MSE가 낮게 나타남
  2. 관측 데이터 셋에서 훈련과 검증용 셋을 나누기 때문에 훈련에 쓰이는 데이터 개수가 줄어들어 성능이 저하됨
     - 적은 양의 데이터로 학습하여 전체 데이터셋으로 학습했을 때보다 에러가 더 클 수 있음 → 테스트 오류율을 과대추정



5.1.2 하나 빼고 교차검증 (LOOCV, Leave-one-out cross-validation)



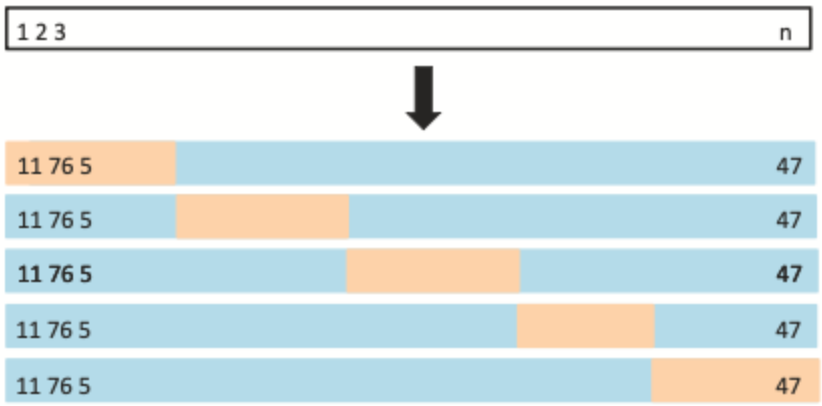
- 단 하나의 관측값을 검증 세트로 사용하고 나머지 관측을 훈련 세트로 사용
- $\mathrm{CV}(n) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$
- 장점
  1. 편향이 적음 (2번 한계 극복)
     - n-1개의 데이터로 훈련하고, 훈련에 사용되지 않은 데이터로 검증하기 때문에 전체 데이터셋으로 훈련하는 것과 성능 유사하면서 테스트 데이터로 검증한 값과 유사한 추정

량을 가짐

2. 반복해도 동일한 결과 반환
    - 무작위 분할이 아닌 고정된 모든 관측값을 검증 데이터로 사용하여 값을 예측하는 것이므로 같은 결과 반환

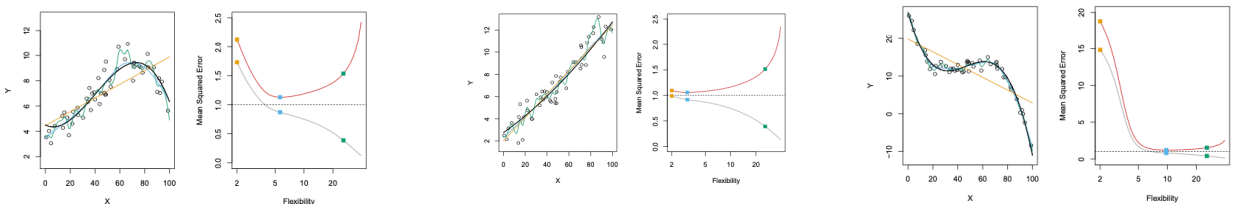
- 단점
    - 하나의 관측에 기반하기 때문에 변동이 심함
    - n번 모델 훈련이 필요하므로 계산량 많음
        - 단, 선형회귀의 경우 단 한번의 모델 훈련으로 LOOCV 값 계산 가능
        - $\mathrm{CV}(n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$
        - h_i : leverage 값. 관측치가 자기 자신의 예측값에 얼마나 영향을 미쳤는지를 나타냄
        - 레버리지가 큰, 모델에 영향을 크게 주는 데이터 포인트가 빠졌을 때 모델 성능이 크게 달라질 수 있으므로 이를 보정하기 위해 레버리지가 클 때 잔차를 좀 더 크게 보도록 1-h로 나눠줌.

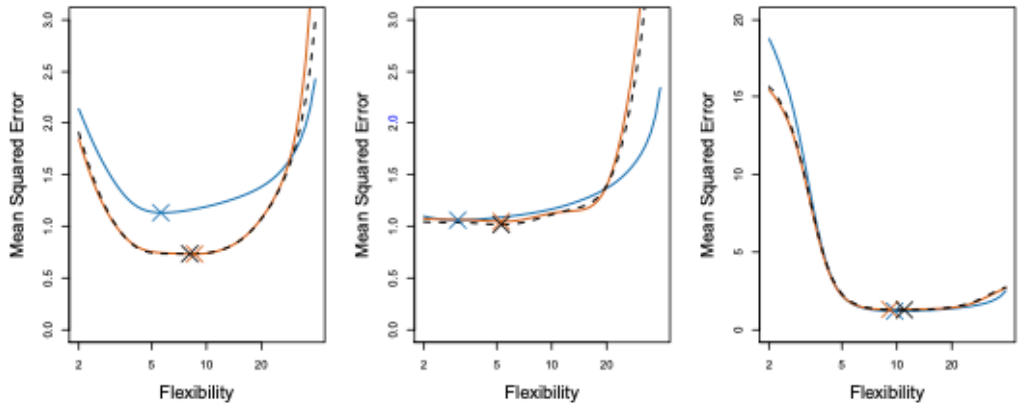
5.1.3 k-겹 교차검증



- 동일한 크기의 k개 그룹으로 무작위 나눠 분할된 집합은 검증 세트로, 나머지 k-1개 집합은 훈련 세트로 사용하며, 이 절차를 매번 다른 관측값 그룹을 검증 세트로 두어 k번 반복
- $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$

- 일반적으로 k=5, 10 사용
- 장점
  - 거의 모든 통계적 학습 방법에 적용 가능
  - 계산 비용 합리적
- 교차검증의 목적
  - 교차검증의 최종 목적은 성능 좋은 모델을 선택하는 것
  - 교차검증의 추정량이 실제 테스트 오차를 얼마나 잘 추정하냐 보다는 가장 성능 좋은 모델을 탐색하는 것이 중요하므로 CV 추정치의 최소값의 위치가 중요
  - CV가 진짜 MSE(파란색 ground truth)를 과대/과소 추정하고 있지만 최소값이 발생하는 위치는 거의 정확히 찾아냄





5.1.4 k-겹 교차검증의 편향-분산 트레이드오프

- **Bias**: 추정값이 실제 값과 얼마나 멀리 있는지
- **Variance**: 샘플에 따라 추정값이 얼마나 흔들리는지

- LOOCV
  - 거의 전체 데이터로 학습 → 편향 작음
  - 비슷한 학습 데이터 사용 → 상관관계 큼 → 높은 분산
- k-fold
  - 학습 데이터가 상대적으로 적으므로 약간의 편향 발생
  - 각 fold 간 학습 데이터셋이 다름 → 모델들 간 상관성 낮음 → 낮은 분산

항목	LOOCV (k=n)	k-Fold CV (예: k=5, k=10)
Bias (편향)	매우 낮음	중간 수준
Variance (분산)	높음	낮음
계산 속도	느림	빠름
추천	이론적으로는 편향이 작지만, 분산 때문에 실제 예측력은 떨어질 수 있음	실제로는 **k=5 또는 k=10**이 가장 안정적이고 널리 사용됨

5.1.5 분류 문제에서의 교차검증

- 회귀 : MSE(평균제곱오차)
- 분류 : 오분류율
- $\mathrm{CV}(n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i, \quad \mathrm{Err}_i = I(y_i \neq \hat{y}_i)$
- 너무 단순하면 언더피팅, 모델이 복잡해질수록 과적합으로 인해 검증 오류율 증가
  - 로지스틱 회귀 적용

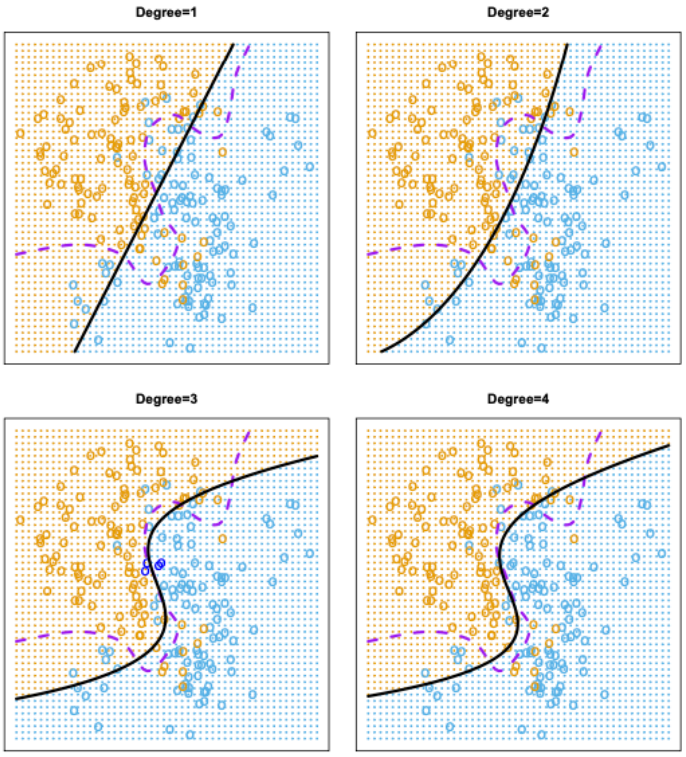
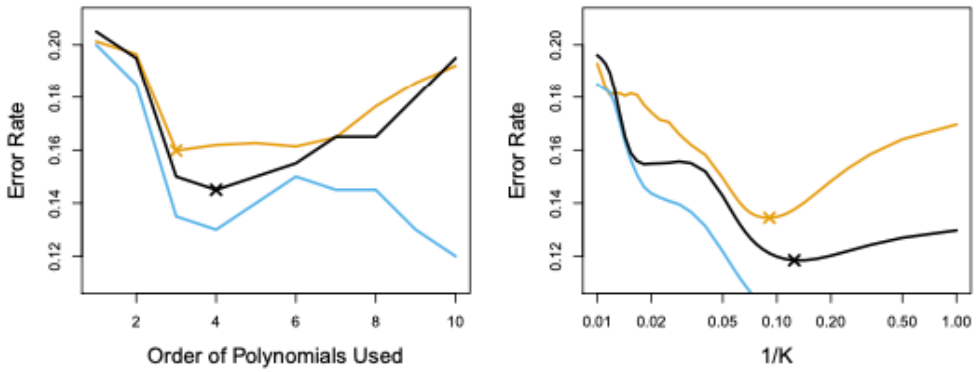
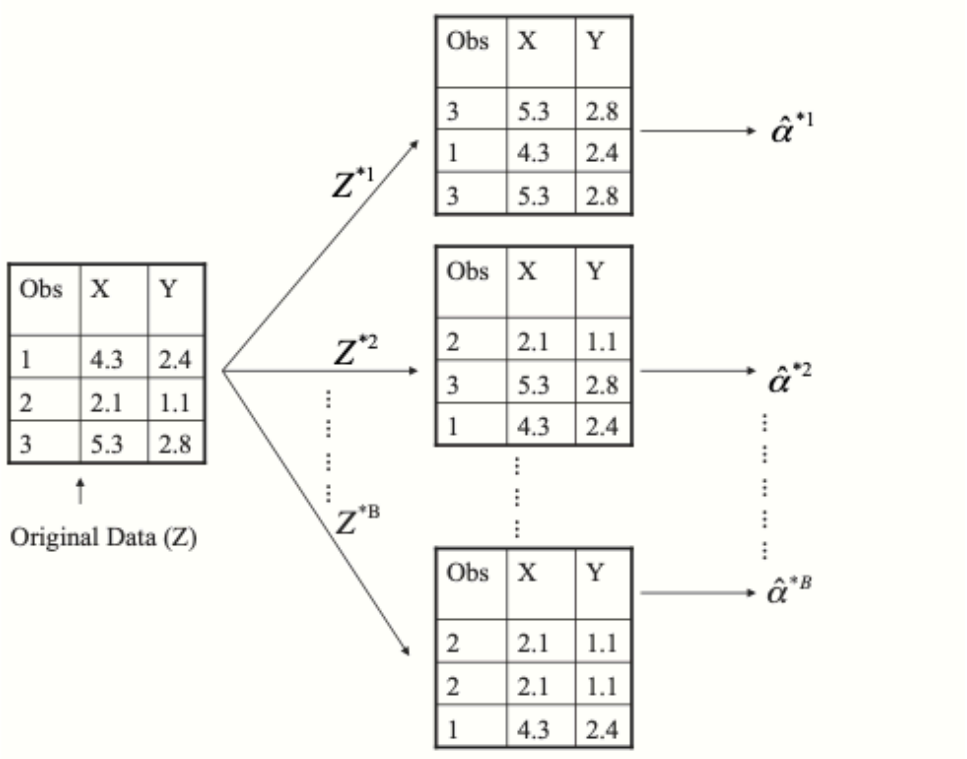


FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively* 0.201, 0.197, 0.160, *and* 0.162, *while the Bayes error rate is* 0.133.

    - 1차 (선형): 0.201
    - 2차 (이차): 0.197
    - 3차 (삼차): 0.160
    - 4차 (사차): 0.162
- CV 오류가 최소가 되는 모델(4차)이 실제 테스트 오류 최소(3차)와 거의 비슷함



5.2 부트스트랩



- 추정치의 불확실성을 정량화할 때 사용
- 선형회귀에서 계수의 표준오차를 추정
- 예) 투자 분산 최소화
    - 두 자산 X, Y의 수익률이 있고, 전체 자산 중 α만큼 X에 투자, 나머지(1-α)를 Y에 투자한다고 가정.

- 전체 포트폴리오의 분산(위험)을 최소화하고 싶은 상황
  - $\min_\alpha \mathrm{Var}(\alpha X + (1-\alpha)Y)$
- 해당 분산을 최소화하는 최적 투자 비율
  - $\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$, ($\sigma = Cov(X,Y)$)
  - 과거 데이터 기반으로 투자 비율 알파 추정
  - $\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$
  - 추정치 알파의 정확도를 알기 위해 표준 오차 추정
  - 이상적인 방법 : 모집단에서 새로운 표본을 추출해 각각 알파를 계산한 후 분산을 구해야 하지만 새로운 표본을 추출하긴 현실적으로 불가능.
- 부트스트랩
  - 모집단에서 새로운 샘플 추출 대신 기존 데이터셋에서 복원 추출로 resampling하여 가짜 데이터셋을 만들기
  - 각 부트스트랩 샘플에서 알파 계산 → 여러 알파의 표준 편차로 SE(표준오차) 추정
  - $\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}_r^* - \bar{\alpha}^*\right)^2}$ ($\hat{\alpha}_r^*$: r번째 부트스트랩 샘플로 얻은 알파 추정치)

❓ 연습문제 풀이

5장 연습문제