Ch.5 재표본추출 방법

- ***재표본추출(Resampling)***이란?
  - 훈련셋에서 반복적으로 표본을 추출하고,
  - 각 표본에 관심있는 모델을 다시 적합하여,
  - 적합된 모델에 대해 추가적인 정보를 얻는 것
- 이러한 접근방식은 모델 적합을 한 번만 하는 경우 얻을 수 없는 정보를 얻을 수 있다.
- 대표적으로 ***교차검증(cross-validation)***과 ***붓스트랩(bootstrap)***이 있다.
  - 교차검증은 모델의 성능을 평가하거나 유연성 수준을 선택하는 데 사용된다.
    - 모델평가(model assessment): 검정오차를 추정하여 모델의 성능을 평가
    - 모델선택(model selection): 적절한 수준의 유연성을 선택
  - 붓스트랩은 모수 또는 통계학습방법의 정확도를 측정하는 데 사용된다.

5.1 교차검증 (Cross-Validation)

- 검정오차율(test error rate)와 훈련오차율(training error rate)
- 2장에서 둘의 차이에 대해 알아보았다.

> 2.2.1 적합의 품질 측정 (p30)

  - 주어진 자료에 대한 통계학습방법의 성능을 평가하고자 할 때, 일반적으로 MSE를 사용한다.
  - MSE(Mean Squared Error) 또는 평균제곱오차는 예측값이 실제값과 얼마나 다른지를 나타낸다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

  - 위는 훈련 데이터를 사용하여 계산되므로, 정확하게 말해서 ***훈련 MSE***라고 한다.
  - 사실 우리가 관심이 있는 것은 통계학습방법을 새로운 데이터에 적용할 때 얻는 예측 정확도이다.
  - 따라서 다음의 ***검정 MSE***가 가능한 작은 모델을 선택하고자 한다.

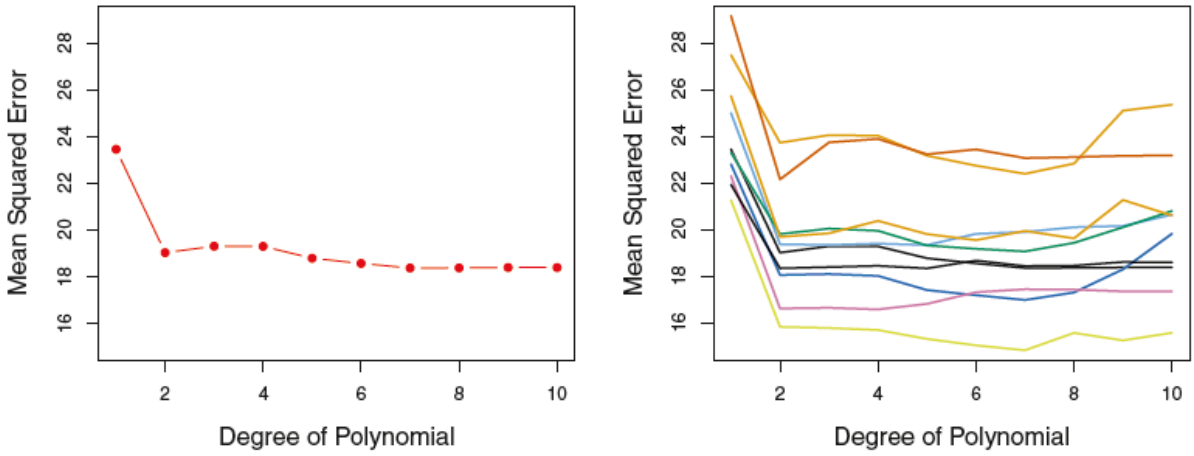
$$Ave[(y_0 - \hat{f}(x_0))^2]$$

- 검정오차는 실제로 관심이 있는 측도이지만, 보통 지정된 검정셋이 없어 계산할 수 없다.
- 훈련오차는 훈련 데이터를 이용하여 쉽게 계산할 수 있지만, 검정오차를 과소추정하는 경향이 있다.
- 따라서, 5장에서는 훈련 데이터를 이용하여 검정오차율을 추정하는 기법들에 대해 알아본다.
  - 5.1.1 ~ 5.1.4: 양적 반응변수, 회귀 문제
  - 5.1.5: 질적 반응변수, 분류 문제

5.1.1 검증셋 기법 (Validation Set Approach)

- 관측치셋을 임의로 두 부분, 훈련셋(training set)과 검증셋(validation set or hold-out set)으로 나눈다.
- 두 부분 중 먼저 *훈련셋*으로 모델을 적합하고, *검증셋*으로 적합한 모델을 예측 및 평가한다.

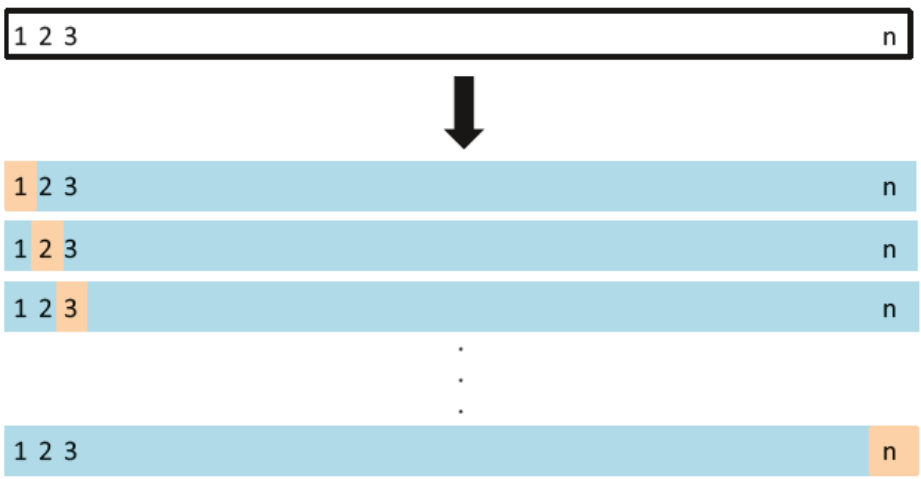
- 즉, ***검증셋 오차율***이 ***검정오차율***에 대한 추정치를 제공한다.
- **<예시>** (그림 5.2)



  - 왼쪽 패널: 2차항 이후로는 검정오차율이 개선되지 않는다.
  - 오른쪽 패널: 10개의 곡선은 서로 일치하지는 않지만, 그러한 경향성이 일관되게 발견된다.
- **<특징>**
-
  - 장점: 개념적으로 단순하고, 구현하기 쉽다.
  - 한계 1: 관측치셋 분할에 따라 검정오차율에 대한 추정치가 크게 변동한다.
  - 한계 2: 훈련셋만을 가지고 모델을 적합해 검정오차율을 과대추정한다.

5.1.2 LOOCV (Leave-One-Out Cross-Validation)

- 검증셋 기법과 비슷하지만 그 한계를 해결하고자 한다.
- 관측치셋을 둘로 분할하는데, 비슷한 크기가 아니라 *하나의 관측치와 나머지 관측치*들로 구성한다.
  - Leave-One-Out, 즉 하나만을 남기고 나머지는 모두 훈련셋으로 구성한다.
  - 하나의 관측치 (x_1, y_1)가 검증셋으로, 나머지 $(x_2, y_2), \ldots, (x_n, y_n)$가 훈련셋으로 사용된다.



- 검정 MSE에 대한 LOOCV 추정치는 ***n개 검정오차 추정치들의 평균***이다.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - $\hat{y}_1$는 x_1을 이용한 예측값으로, 이때 (x_1, y_1)은 모델적합에 사용되지 않았다.
  - 따라서 $MSE_1 = (y_1 - \hat{y}_1)^2$는 검정오차에 대한 *approximately unbiased* 추정치를 제공한다.

    - 그러나 하나의 관측치 (x_1, y_1)에 기초하므로 변동이 커서 좋지 않은 추정치이다.
        - Desirable properties of estimator: 1) consistency 2) unbiasedness 3) efficiency
        - Generally MLE is preferred as it satisfies 1) consistency and 3) efficiency (smaller variance), which are baselines, and also invariance property (and asymptotically unbiased).
    - 따라서 $MSE_1, \ldots, MSE_n$까지 n번 반복하여 평균을 낸다.
- **<특징>**
    - 장점 1: (검증셋 기법과 비교하여) 편향이 작다.
        - n-1개 관측치를 포함하는 훈련셋으로 적합하므로 검정오차율을 과대추정하지 않는다.
    - 장점 2: (검증셋 기법과 비교하여) 관측치셋 분할과 관계없이 항상 같은 값을 얻는다.
    - 장점 3: 매우 일반적인 방법이고, 어떠한 종류의 모델과도 사용될 수 있다.
    - 한계: 모델을 n번 적합해야 하므로 구현 부담이 있다.
- 일반적으로 성립하는 것은 아니지만, ***레버리지(leverage)***를 이용해 계산시간을 단축시킬 수 있다.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{1-h_i})^2$$

    - 레버리지 h_i는 $1/n$과 1 사이에 놓이고, 어떤 관측치가 적합에 주는 영향을 의미한다.

> 3.3.3 잠재적 문제 (p109)

        - 이상치는 주어진 설명변수의 값 x_i에 대해 반응변수의 값 y_i가 보통 수준과 다른 것이다.
        - 반면 높은 레버리지를 갖는 관측치는 x_i값이 보통 수준과 다르다.
        - 이상치와 달리 높은 레버리지 관측치는 추정회귀선에 상당한 영향을 주기 때문에 식별이 중요하다.
    - 따라서 레버리지 h_i가 높으면 (= 분모인 $1-h_i$가 작으면) MSE^*가 부풀려진다.
    - 일반 LOOCV가 n번 적합시키는 것과 달리, h_i를 포함시킨 하나의 모델만을 적합하여 시간을 단축한다.

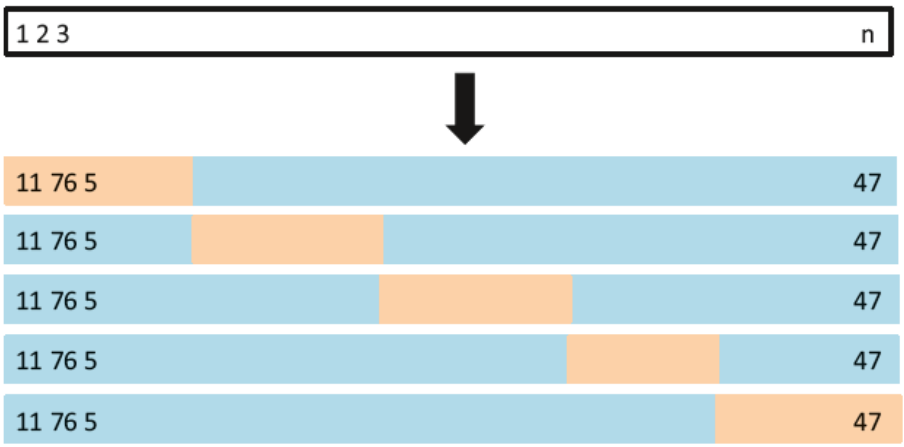
5.1.3 k-fold 교차검증

- 관측치셋을 임의로 *크기가 비슷한 k개 그룹 (또는 fold)*로 분할한다.
- 첫 번째 fold는 검증셋으로, 나머지 (k-1)개 fold를 훈련셋으로 사용하며 이러한 과정을 k번 반복한다.
- 즉 검정 MSE에 대한 k-fold CV 추정치는 ***k개 검정오차 추정치들의 평균***이다.

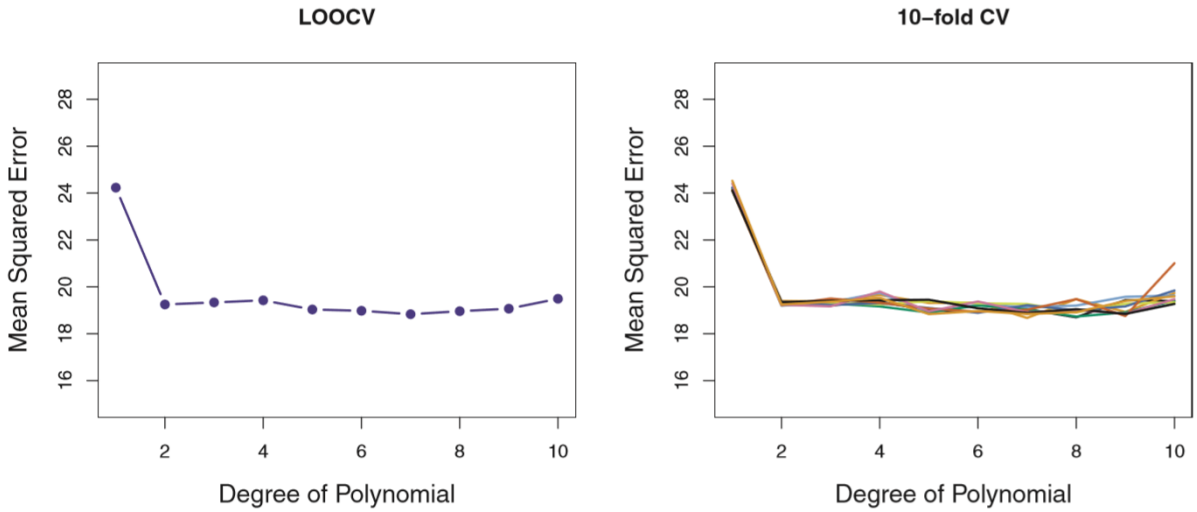
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

    - LOOCV는 k를 n과 동일하게 설정한 k-fold CV의 특별한 경우다.
    - real data에는 보통 k=5 혹은 k=10을 사용한다.

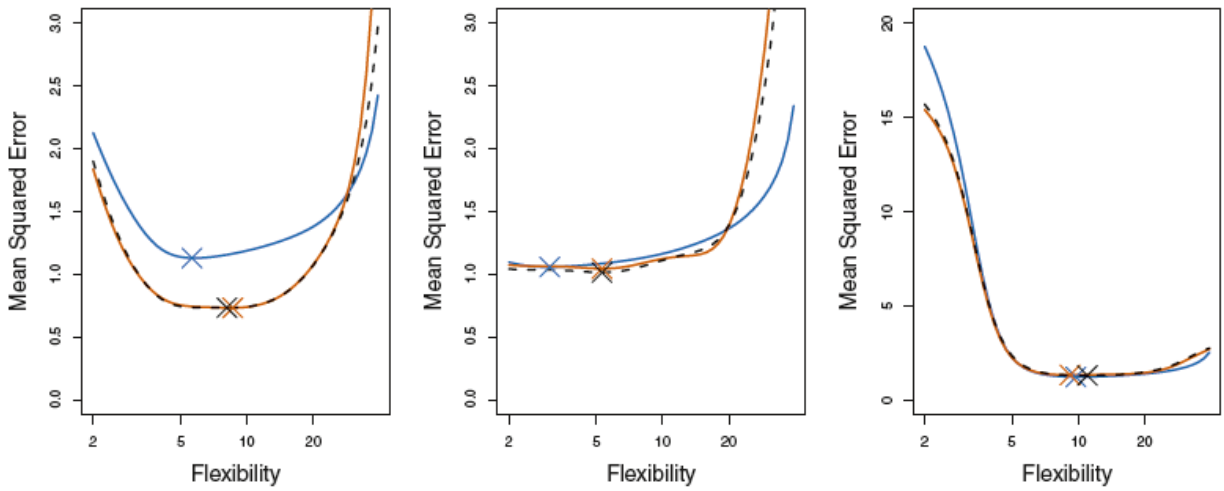
- k=5인 경우, 아래와 같이 5개의 겹치지 않는 그룹으로 분할된다.



- **<예시>** (그림 5.4)



  - 왼쪽 패널: LOOCV 오차곡선
  - 오른쪽 패널: 10-fold CV 오차곡선이 9개로, 그림 5.2에 비해 변동이 훨씬 작다.
- **<특징>**
  - 장점 1: (LOOCV에 비해) 계산량이 n번에서 k번 적합으로 줄었다.
  - 장점 2: (LOOCV에 비해) 편향-분산 절충으로 인해 정확도가 높다.
- 물론 CV 곡선들은 실제 검정 MSE와 다른 값을 추정할 것이다.



  - [파란색: 실제 검정 MSE], [검은색 파선: LOOCV 추정치], [오렌지색: 10-fold CV 추정치]

- 오른쪽 패널은 거의 일치하지만, 중앙과 왼쪽 패널은 과소추정하거나 과대추정한다.
- 때로 교차검증을 수행하는 목적은 최소 MSE값을 갖는 ***위치***를 찾는 것이며, 이때 MSE값은 중요하지 않다.
- 이 경우 CV 곡선은 최소의 검정 MSE에 대응하는 ***유연성 수준***을 식별할 수 있게 한다.

5.1.4 k-fold 교차검증에 대한 편향-분산 절충

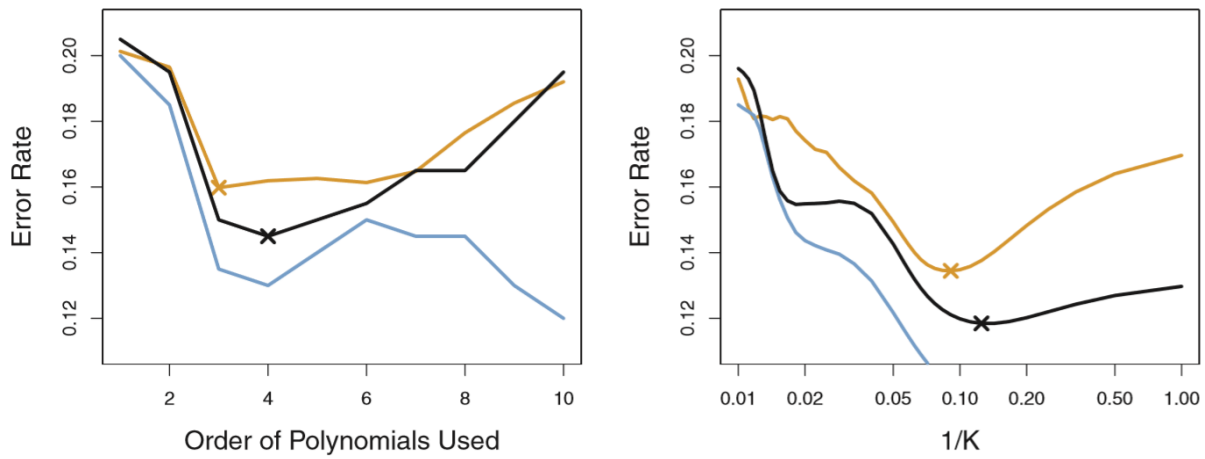
- k-fold 교차검증의 장점 2에 대해 좀 더 알아본다.
  - 장점 2: (LOOCV에 비해) 편향-분산 절충으로 인해 정확도가 높다.
- LOOCV의 편향: 전체 데이터셋과 거의 같은 n-1개 크기의 훈련셋을 이용하므로, 편향이 작다.
- LOOCV의 분산: 그러나 n개 모델은 거의 동일한 관측치들을 사용하므로, 그 결과 역시 (양의) 상관성이 높다.
- k-fold 교차검증은 편향 감소의 측면에서는 LOOCV보다 못하지만, 분산 감소에서는 더 낫다.
  - k-fold 교차검증은 겹치는 부분이 적으므로 상대적으로 상관성이 낮다.
  - 상관성이 높은 값들의 평균은 그렇지 않은 평균보다 분산이 크기 때문에, LOOCV가 분산이 더 크다.
- 요약하면, k값의 선택과 관련된 편향-분산 절충이 존재한다.
  - k가 클수록 편향이 줄어들지만, 분산 역시 증가하여 정확도가 떨어진다.
  - k=5 혹은 k=10인 경우 적절하다는 것이 알려져 있다.

5.1.5 분류문제에 대한 교차검증

- 분류문제에서는 MSE 대신 ***잘못 분류된 관측치의 수***를 사용한다.
- 예를 들어, LOOCV의 오차율은 다음과 같다.

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} Err_i = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- **<예시>**



  - 유연성을 증가시키고자 로지스틱 회귀모델에 다항식을 추가한다고 하자.
  - [갈색: 검정오차], [파란색: 훈련오차], [검은색: 10-fold CV]
  - 왼쪽 패널: 로지스틱 회귀에서 10-fold CV 오차율은 검정오차율에 잘 근사된다.
  - 오른쪽 패널: k값에 따른 KNN 분류기로, 훈련오차율은 유연해질수록 감소하여 사용할 수 없다.
  - 두 경우 모두 10-fold CV 곡선은 최적의 차원 수, K값을 찾는 데 유용하다.

5.2 붓스트랩 (Bootstrap)

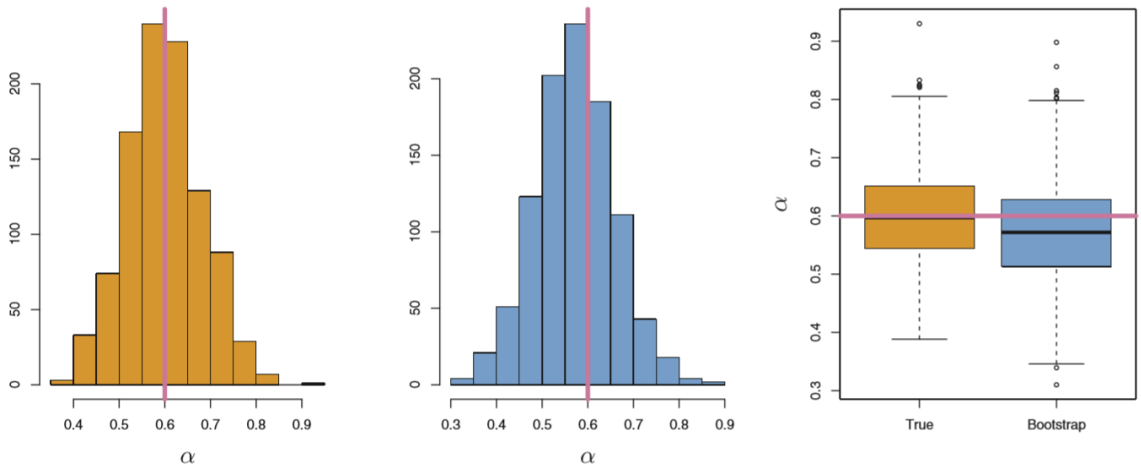
- Bootstrap은 *불확실성*을 수량화하는 데 사용할 수 있는 강력한 통계적 도구이다.
  - 선형회귀는 R에서 자동으로 표준오차를 계산해주지만, 보다 복잡한 기법이라면 불가능할 수 있다.
  - 통계소프트웨어가 변동의 측도를 자동으로 제공해주지 않는 경우 유용하다.
- 시나리오: 임의의 투자수익 X와 Y를 각각 얻을 수 있는 두 가지 금융자산에 일정한 금액을 투자한다. 전체 투자 금액의 비율 α를 X에, $1-\alpha$를 Y에 투자할 것이다. 투자의 전체 위험 또는 분산을 최소화하도록 α를 선택한다.
  - 즉, $Var(\alpha X + (1-\alpha)Y)$를 최소화하고자 하며, 그 α는 다음과 같이 구한다.

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

  - 모의 투자수익을 사용하여 $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$를 추정하고, 추정값들을 위 식에 대입하여 $\hat{\alpha}$를얻는다.
  - $\hat{\alpha}$에 대한 정확도를 수량화해보자.
    - α를 추정하는 과정을 1,000번 반복하면 $SE(\hat{\alpha}) \approx 0.083$를 구할 수 있다.
    - 모집단의 랜덤표본에 대해 $\hat{\alpha}$는 α와 평균적으로 0/083만큼 다를 것으로 기대된다.
    - 그러나 실제로는 모집단으로부터 새로운 표본들을 생성할 수 없으므로, 위는 불가능한 방법이다.
  - Bootstrap 기법은 추가로 표본을 생성하지 않고 $\hat{\alpha}$의 변동을 추정한다.
- Bootstrap 기법은 컴퓨터를 이용해 새로운 표본 셋을 얻는 과정을 모방한다.
- 모집단에서 독립적인 데이터셋들을 얻는 대신, 원래의 데이터셋으로부터 관측치를 반복적으로 추출한다.
  - 원래의 데이터셋에서 n개의 관측치를 랜덤하게 추출하여 Bootstrap 데이터셋 Z^{*1}를 얻는다.
  - 이 때 표본추출은 복원추출 (with replacement) 방식으로, 동일한 관측치가 포함될 수 있다.
  - Z^{*1}를 사용하여 $\hat{\alpha}^{*1}$에 대한 추정치를 얻는다.
  - 이 절차를 B번 반복하여 B개의 Z^*와, B개의 $\hat{\alpha}^*$를 얻는다.
  - 그러한 Bootstrap 추정치들의 표준오차는 다음 식으로 계산된다.

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}. \qquad (5.8)$$

  - 히스토그램과 박스플롯을 그렸을 때 실제 [orange] 와 Bootstrap [Blue] 추정치가 상당히 비슷하다.



Ch.6 선형모델 선택 및 Regularization

- 아래의 표준적인 선형모델은 보통 최소제곱을 사용하여 적합한다.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

  - 7, 8장에서 선형모델을 확장시키기 전에, 단순선형모델을 개선할 수 있는 방법을 알아본다.
  - 선형모델은 추론(해석)의 관점에서, 또한 현실적 측면에서 분명한 경쟁력이 있다.
  - 따라서 최소제곱적합을 다른 적합절차로 대체하여 개선해보고자 한다.
- 최소제곱 대신 다른 적합절차를 사용하는 이유는?
  1. 예측 정확도
     - 관측치의 수 n >> 변수의 수 p라면 최소제곱 추정치의 분산이 커져 정확도가 떨어진다.
     - 심지어 n < p라면 분산이 무한대가 되어 최소제곱법을 사용할 수 없다.
     - 추정하는 모수의 수를 ***제한(constrain)*** 또는 ***수축(shrink)***한다면, ignorable한 편향 증가를 희생하여 분산을 현저하게 감소시킬 수 있으며, 곧 예측 정확도가 향상된다.
  2. 모델 해석력
     - 관련이 없는 변수들을 제외하여 좀 더 해석하기 쉬운 모델을 얻을 수 있다.
     - 최소제곱법으로는 0인 계수 추정치를 얻게 될 가능성은 거의 없다.
     - ***특징선택(feature selection)*** 혹은 ***변수선택(variable selection)***을 수행해보자.
- 최소제곱 대신 사용한 적합 방법으로는 다음의 세 가지를 알아본다.
  1. 부분집합 선택
     - 반응변수와 관련이 있다고 생각되는 부분집합을 식별하여 최소제곱을 사용한다.
  2. 수축 (shrinkage 또는 regularization)
     - p개의 설명변수를 모두 포함하지만 일부 0으로 수축된다.
  3. 차원축소 (Dimension Reduction)
     - p개의 설명변수를 M차원 (M < p) 부분공간으로 투영 (projection) 시킨다.
     - 그런 다음 M개의 투영은 최소제곱에 의해 선형모델의 설명변수로 사용된다.

6.1 부분집합 선택

6.1.1 최상의 부분집합 선택

- p개 설명변수의 all possible 조합 각각에 대해 최소제곱회귀를 적합한다.
- 즉 설명변수가 하나인 모델, 2개인 모델, ..., p개인 모델 중 최고의 모델을 선택한다.

- **<알고리즘 6.1>**

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

  - Step 1. 먼저 영모델로 시작한다.
  - Step 2. (a) 2^p의 가능한 모델에서 하나를 선택하는 문제에서 (b) (p+1) 개 중 하나로 축소한다.
    - 각 조합에서 최고 모델을 선택할 때에는 RSS나 R^2를 사용한다.
  - Step 3. RSS나 R^2의 대안으로 교차검증된 예측오차나 $C_p, AIC, BIC, adjusted\ R^2$를 사용한다.
    - 우리가 선택하고자 하는 것은 훈련오차가 아니라 검정오차가 낮은 모델이다.
    - RSS나 R^2는 변수의 수가 늘어남에 따라 단조감소 / 증가하므로 주의한다.
- 로지스틱 회귀의 경우 RSS 대신 ***이탈도 (deviance)***를 사용한다.
  - 이탈도란 최대 로그우도 (maximized log-likelihood)를 -2배한 것이며, 그 값이 작을수록 잘 적합된다.
- 간단하고 개념적으로 흥미로운 기법이지만 계산상의 제약이 있다.
  - p = 20이면 100만 개 이상의 모델이 가능하다.
  - 계산량을 줄이기 위해 일부 선택을 제외하는 *분기한정기법 (branch-and-bound technique)*도 있다.
  - 그러나 이 기법 역시 p값 증가에 따른 제약이 있고, 최소제곱 선형회귀에만 적용된다.

6.1.2 단계적 선택

- 최상의 부분집합 선택은 계산상의 이유로 p가 아주 크면 사용할 수 없다.
- 또한 검색공간이 거대하면 잘 맞는 모델을 찾을 가능성이 높지만, 과적합이나 높은 분산을 유발할 수 있다.
- 따라서 훨씬 제한된 모델들의 집합을 조사하는 *단계적 방법*이 대안이 된다.

> ***1. 전진 단계적 선택***

- 최상의 부분집합 선택에 비해 훨씬 적은 수의 모델들을 고려한다.
- 각 단계에서 적합이 가장 크게 향상되는 변수를 모델에 추가한다.

- **<알고리즘 6.2>**

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

  - Step 1. 먼저 영모델로 시작한다.
  - Step 2. k (k = 0, ..., p-1) 번째 반복에서 남은 변수의 수는 (p-k)개이므로, (p-k)개 모델을 적합한다.
    - 총 $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ 개의 모델을 적합한다.
  - Step 3. 마찬가지로 서로 다른 변수의 수를 가진 모델들을 비교한다.
- 최상의 부분집합 선택에 비해 계산상 장점이 분명하지만, 최고의 모델을 찾는다는 보장은 없다.
- n < p인 고차원 설정에서도 적용할 수 있지만, 부분모델 $M_0, \ldots, M_{n-1}$만 가능하다.

> *2. 후진 단계적 선택*

- p개의 설명변수 모두를 포함하는 모델에서 시작해, 유용성이 작은 변수를 반복적으로 제외시킨다.
- **<알고리즘 6.3>**

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

  - Step 1. Full model에서 시작한다.
  - Step 2. 하나의 설명변수를 제외한 모든 k개의 모델을 고려하여 최고 모델을 선택한다.
  - Step 3. 마찬가지로 서로 다른 변수의 수를 가진 모델들을 비교한다.
- 최상의 부분집합 선택에 비해 계산상 장점이 있지만, 최고의 모델을 찾는다는 보장은 없다.
- 전진 단계적 선택법과 달리 n > p인 경우에만 사용할 수 있다.

> ***3. 하이브리드 방식***

- 전진 단계적 선택과 후진 단계적 선택의 하이브리드 버전이 있다.
- 둘의 계산적 장점을 유지하면서 최상의 부분집합 선택을 모방하고자 한다.

6.1.3 최적의 모델 선택

- 우리는 훈련오차가 아니라 검정오차가 가장 낮은 모델을 선택하고자 한다.
- Step 3에서는 변수의 수가 다른 모델을 비교하므로 RSS나 R^2는 좋지 않다.
- 검정오차를 추정하는 두 가지 기법은 다음과 같다.
    1. 간접 추정: 과적합으로 인한 편향을 고려하도록 훈련오차를 조정한다.
    2. 직접 추정: 검증셋 기법 또는 교차검증 기법으로 직접 추정한다.

1. 간접 추정

- 모델의 크기에 따라 훈련오차를 조정하는 기법을 사용할 수 있다.
- 즉, 다음의 네 가지 기법에는 변수의 수인 d가 고려된다.

> ***1. (Mallow's)*** C_p

- 설명변수가 d개인 최소제곱 모델에 대해 C_p 추정치는 다음과 같다.

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

    - 이때 $\hat{\sigma}^2$는 오차 ϵ의 분산에 대한 추정치이다.
    - 즉 설명변수의 수가 증가할수록 훈련 RSS에 페널티를 더욱 많이 부과한다.
    - $\hat{\sigma}^2$가 σ^2의 unbiased estimator라면 C_p는 검정 MSE의 unbiased estimator임이 알려져 있다.
    - 결론적으로, C_p 통계량은 낮은 검정오차를 갖는 모델에 대해 작은 값을 갖는다.
    - 그렇다면 최고 모델을 결정할 때 가장 낮은 C_p 값을 가지는 모델을 선택하면 된다.

> ***2.*** AIC ***(Akaike information criterion)***

- 최대 가능도에 의해 적합된 모델들로 구성된 하나의 클래스에 대해 정의된다.
- 오차들이 정규분포를 따를 경우 최대 가능도와 최소제곱은 같다. (C_p와 AIC도 같다.)
- 추가적인 상수를 생략한 AIC는 다음과 같다.

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

    - 결과적으로 최소제곱 모델의 경우 C_p와 AIC는 비례한다.

> ***3.*** BIC ***(Bayesian information criterion)***

- 베이즈 관점에서 파생되었지만 C_p 및 AIC와도 유사하다.

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

    - BIC는 C_p에서 2를 $log(n)$으로 대체함으로써, 변수의 수가 크면 더 많은 페널티를 부과한다.
    - 그 결과 C_p보다 더 작은 크기의 모델이 선택된다.

> ***4. Adjusted*** R^2

- 설명변수의 수가 늘어남에 따라 RSS는 항상 감소하고 R^2는 항상 증가한다.

- 따라서 설명변수의 개수 d를 고려한 R^2 통계량은 다음과 같다.

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

  - 앞의 세 경우와 달리, 조정된 R^2는 값이 클수록 모델의 검정오차가 작다는 것을 의미한다.
  - 노이즈(noise) 변수를 추가하는 것은 d를 증가시키므로, $RSS/(n-d-1)$를 증가시킬 것이다.
  - 즉 불필요한 변수들을 포함하는 것에 대한 대가를 지불한다.

2. 직접 추정

- Ch. 5 검증 및 교차검증 방법을 사용하여 검정오차를 직접 추정할 수 있다.
- 간접 추정과 달리 직접적인 추정치를 제공하고, 실제 모델에 대한 가정이 적다는 장점이 있다.
- 또한 더 넓은 범위에 걸쳐 사용될 수 있으며, 모델의 자유도나 오차항 분산을 모르더라도 가능하다.
- 과거에는 계산상 제약으로 간접 추정이 더 주목받았으나, 최근 컴퓨터 기술의 발달로 아주 유용하다.
- 관측치셋 분할이나 fold 크기를 달리 하면 *one-standard error* 규칙을 사용한다.
  - 먼저 각 모델 크기에 대해 추정된 검정 MSE의 표준 오차를 계산한다.
  - 그 다음 검정 MSE 곡선에서 가장 작은 값의 1-표준오차 이내에 있는 추정된 검정오차를 확인한다.
  - 마지막으로 검정오차가 해당 범위에 있는 모델들 중 가장 크기가 작은 것을 선택한다.
  - 근거는 모델이 비슷한 수준이라면 더 단순한 모델을 선택하고자 함이다.