5. Resampling Methods

Resampling methods는 현대 통계학에서 매우 중요한 도구이다. training set에서 샘플을 뽑아서 모델에 적합시키는 과정을 반복하게 된다. 이번 장에서는 cross-validation과 bootstrap에 대해서 알아볼 것이다. 먼저 cross-validation은 test-error rate을 측정하거나(model assessment) 적절한 flexibility를 찾는 과정(model selection)에 사용된다. bootstrap은 파라미터 추정치나 통계 모델의 정확도를 측정하기 위해서 주로 사용된다.

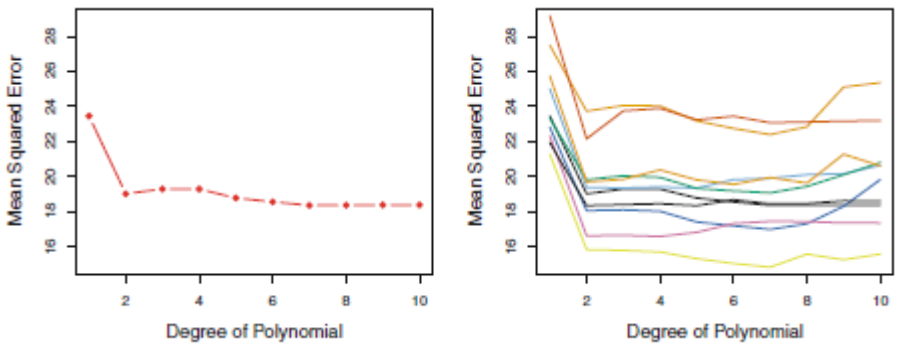
5.1 Cross-Validation

training error rate은 구하기 쉬운 반면에, test error rate은 구하기 어렵다. 왜냐하면 지정된 test set이 없기 때문이다. 그래서 이번 장에서는 training observation에서 subset을 추출하여 test error rate을 구하는 방법에 대해서 알아볼 것이다.

5.1.1 The Validation Set Approach

validation set approach는 간단하다. observation을 training set과 validation set 2개로 나눈다. 모델은 training set을 이용해서 적합을 하고 나머지 validation set을 통해 적합된 모델의 test error rate을 추정한다.

그러나 이러한 방법에는 문제점이 있다. 이 문제점을 확인하기 위해 validation set approach를 여러 번 반복해보았다.



이 실험을 통해 알 수 있는 점은 선형적인 것보다는 2차이상인 것의 MSE가 더 낮다, 통일된 MSE를 알 수 없다는 점이다.

즉, validation set approach는 다음의 2가지 문제점이 있다.

1. variance가 매우 크다.
2. 이 approach를 할 때 쓰이는 데이터의 수가 적기 때문에, test error rate을 overestimate 할 수 있다.

따라서 다음부터 나올 기법들은 이러한 문제점을 해결하는 기법들이다.

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

LOOCV는 1개의 validation set만을 사용하는 것이다. 이렇게 총 n번의 과정을 반복하고 그것들의 평균을 낸다. 그러면 통일된 MSE가 나오고, 모든 데이터를 다 사용해서 overestimate하는 문제를 완화할 수 있다.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

이때, LOOCV는 2개의 장점을 가진다. 첫번째는 bias가 작다는 것이다. 이는 모든 데이터를 다 사용했기 때문이다. 두번째로는 통일된 mse가 나온다는 점이다.

그런데, 하나하나 다 하기 때문에, n값이 매우 큰 경우에는 시간이 많이 걸리고 비용이 큰 문제점이 발생한다. 그래서 이러한 문제를 해결하기 위해 least square linear나 polynomial regression을 통해 LOOCV의 비용을 줄인다.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

원래의 MSE값에 i번째 잔차를 1-h로 나눈것만 다르다. 이때 h는 leverage이다. (180쪽의 hence the residuals for high-leverage points are inflated in this formula by exactly the right amount for this equality to hold는 뭔소리지?)

5.1.3 k-Fold Cross-Validation

k-fold CV는 k개의 그룹으로 나누는 것이다. 만약에 100개의 데이터가 있고 K=5라고 한다면, 20개씩 묶어서, 전에 했던 것과 같이 하는 것이다.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

이 경우에는 LOOCV와는 달리 computational 문제가 덜하다. 왜냐하면 연산횟수가 줄어들기 때문이다. 또한 뒤에서 논의하겠지만, k-fold CV는 bias가 조금 있는 대신에 variance도 크지 않다. 반면에 LOOCV는 unbiased한 대신에 variance가 커서 문제점이 있다.

5.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation

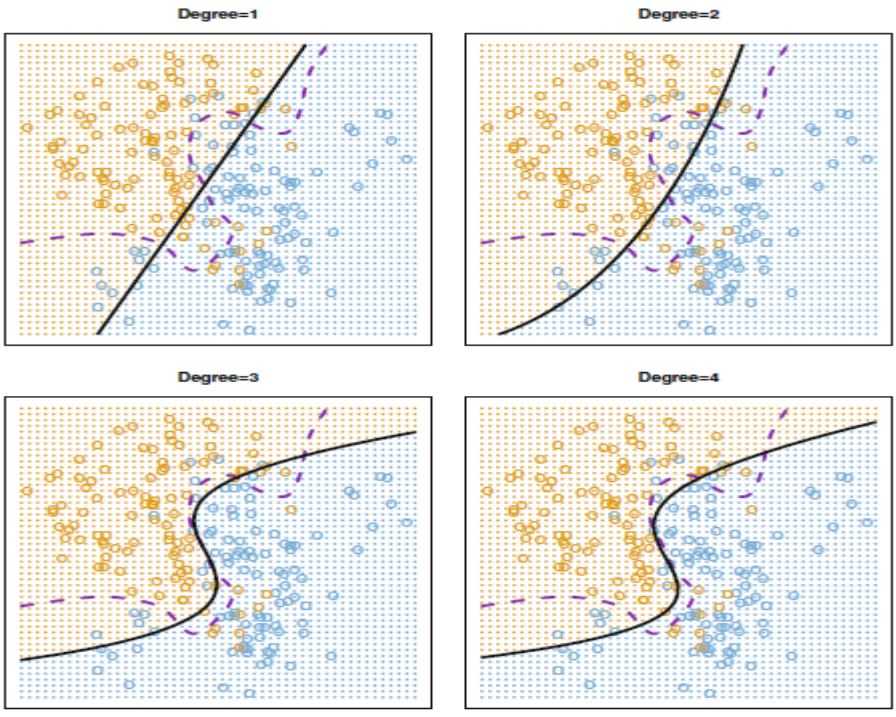
bias의 크기를 놓고 보면, LOOCV < k-fold CV < validation approach이다. 그러나 bias-variance trade-off가 있기 때문에, LOOCV는 variance가 크다고 예측할 수 있다. 왜냐하면, LOOCV를 쓸 때에는 모든 데이터를 하나하나 다 쓰기 때문에 변수들 간에 상관관계가 있을 수 있다. 반면에 k-fold CV는 그럴 가능성이 적어진다. 따라서 k-fold CV가 더 괜찮은 경우가 많다. k값의 경우에는 경험적으로 봤을 때, 5나 10일 때가 가장 적절하다고 한다. bias나 variance값이 지나치게 크지 않기 때문이다.

5.1.5 Cross-Validation on Classification Problems

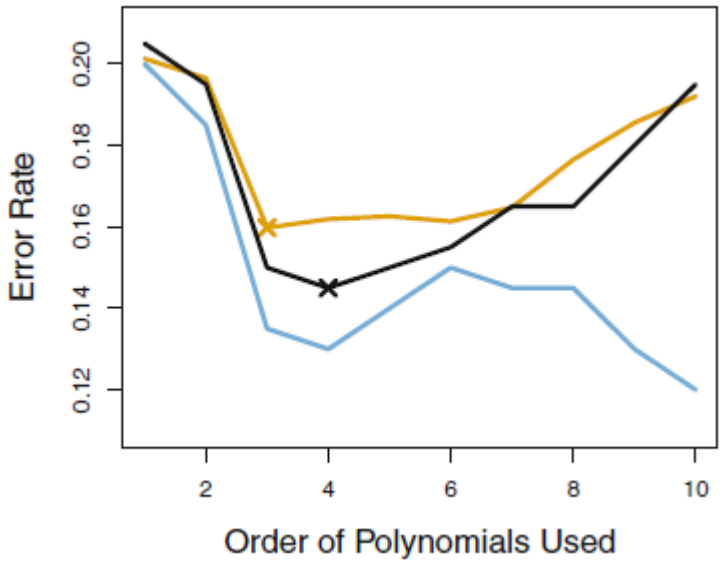
지금까지는 반응변수 Y가 양적일 때를 살펴봤다. 이제부터는 반응변수가 질적인 경우인 classification일 때의 CV를 살펴볼 것이다. MSE대신에 잘못분류된 관측치의 수를 사용할 것이다.

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$

$$\text{where } \mathrm{Err}_i = I(y_i \neq \hat{y}_i).$$



예를 들어, logistic regression을 이용해서 decision boundary를 추정했다고 했다고 해보자. 위의 네 개의 그림은 각각 함수의 차수를 높여본 것이다. 각각의 test error rate은 0.201, 0.197, 0.160, 0.162이다. 보라색 점선이 Bayes decision boundary이다. 그런데, 실제 상황에서는 true test error rate과 Bayes decision boundary를 알 수 없다.



위의 그림은 logistic regression의 차수를 높였을때의 error rate을 나타낸 것이다. 파란색 선은 training error이고 검정색 선은 10-fold CV error, 노란색 선은 test error이다. 모델의 flexibility가 높아질수록, training error는 전체적으로 감소하는 형태를 띠고, test error는 U자 모양을 띤다. 10-fold CV error도 비슷한 모양을 띠고 있다. 비록 CV error가 underestimate하는 경향은 있지만, 4차일 경우에 최저점을 찍는다는 점은 동일하다. 따라서, 앞선 예시에서는 4차의 logistic regression이 Bayes decision boundary에 가장 가깝다는 점을 알 수 있다.

5.2 The Bootstrap

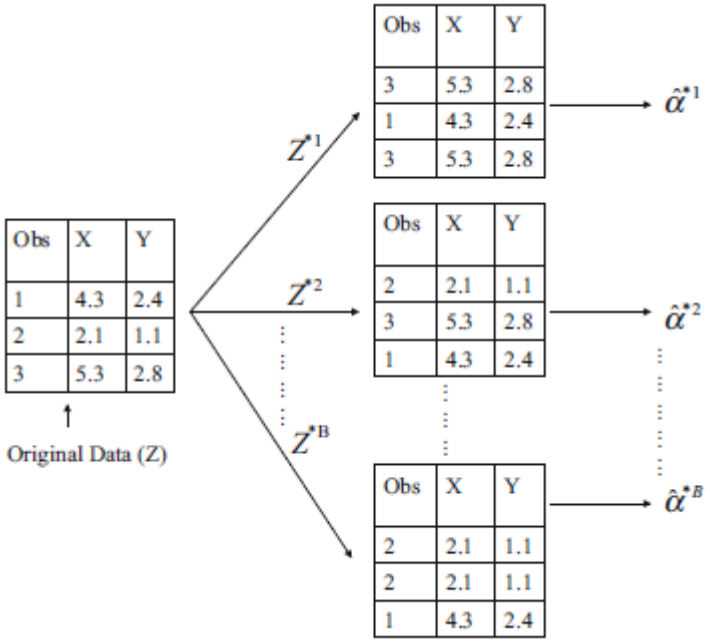
bootstrap은 추정치나 통계기법들에 있는 불확실성을 수치화하는 데에 쓰인다. 예를 들어, 선형회귀에서의 상관계수들의 표준오차를 추정할 때 쓰인다. 그런데 R과 같은 프로그램에서는 이를 자동으로 계산해준다. 따라서 variability(변동성;분산)를 얻기 어렵거나 자동으로 계산되지 않는 경우에 bootstrap이 쓰인다.

예를 들어 우리가 X와 Y에 각각 a, 1-a만큼 투자한다고 하자. 우리는 이때, 투자의 위험성을 최소화해야한다. 즉, Var(aX+(1-a)Y)가 최소가 되는 a값을 찾아야한다. 이는 아래와같다.

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

실제 모수는 알 수 없어서 hat을 붙였다. 그 추정의 과정은, X와 Y의 이익금의 표본 100개를 이용하여 추정하였다. 그 결과 a hat의 값은 0.532와 0.657사이에 있었다. 더 정확히 하기 위해서 표본 1000개를 이용해서 a hat값의 평균을 구해보니 0.5996이었고 표준편차는 0.083이었다. 그리고 a값의 표준편차는 0.08이었다. 이를통해 우리가 추정한 값이 매우 정확한 편이라는 것을 알 수 있다.

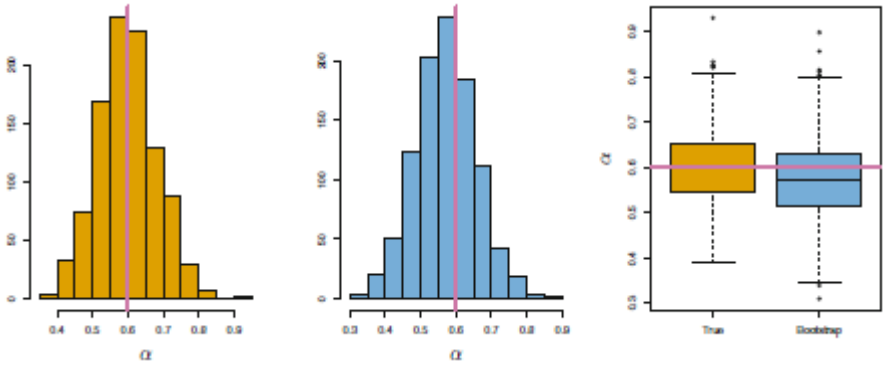
하지만, 우리는 새로운 표본을 더 추출할 수 없기 때문에, 위의 과정은 할 수 없다. 그러나 bootstrap이 이를 가능하게 해준다. 오리지널 데이터로 다른 데이터셋을 더 만드는 방식이다.



예를 들어 n = 3이라고 하자. 즉, 관측치가 3개이다. 그러면 B개 만큼의 데이터셋을 만든다. 이때, 중복이 가능하다. 그리고 각각의 만들어진 B개의 데이터셋으로 a값을 추정한다. 그리고 그것들의 표준편차를 다음의 식으로 구한다.

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

아까의 이상적인 경우와의 예시로 다시 돌아가서 보면 다음과 같다.



왼쪽의 경우에는 새로 1000개의 데이터를 뽑은 것이고, 가운데 그림에서는 1000개의 bootstrap을 이용한 경우이다. 가운데의 경우에 표준편차는 0.087로 좌측의 경우의 0.083 보다는 크지만 0.08과 역시 큰 차이가 없다는 점을 알 수 있다. 이는 bootstrap이 매우 효과적으로 변동성을 측정하는 도구라는 뜻이다.