00 과목 소개

J Kim

2021.3

Outline

교재 소개
수업
과제
평가

교재 소개

- 원서
  - G. Janes, D. Witten, T. Hastie, R. Tibshirani (2013). *An introduction to statistical learning with applications in R*. Springer
  - pdf 파일은 아래 주소에서 다운로드 가능
    http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
- 한글 번역본은 인터넷 서점에서 구매 가능
  - 마이클 (2016). 가볍게 시작하는 통계학습. 루비페이퍼
- 교재에서 다룰 자료 파일은 아래 주소에서 다운로드 가능
  - https://cran.r-project.org/web/packages/ISLR/index.html

목차: 응용통계1

- 1장 Introduction
- 2장 Statistical learning – 용어와 개념, K-nearest neighbor classifier 등
- 3장 Linear regression – 단순회귀모형, 중회귀모형 등
- 4장 Classification – 로지스틱 회귀모형, 판별분석 등
- 5장 Resampling methods – cross-validation, bootstrap 등
- 6장 Linear model selection and regularization – stepwise selection, ridge regression, lasso 등

목차: 응용통계2

- 6장 Linear model selection and regularization – principal component regression, partial least squares 등
- 7장 Moving beyond linearity – non-linear methods
- 8장 Tree-based methods – bagging, boosting, random forests 등
- 9장 Support vector machine – both linear and non-linear methods
- 10장 Unsupervised learning – principal component analysis, K-means clustering, hierarchical clustering 등

수업

- 실시간 줌으로 수업 예정이며 접속 아이디는 https://us02web.zoom.us/j/3700952329
- PPT를 이용한 수업진행
  - 수업 전에 강의록은 canvas 강의자료실에 업로드할 예정
  - 강의록은 각자 출력!
- 강의록에 담지 못한 내용은 판서로 보충할 예정
- 강의와 R 실습을 병행하여 진행
- 1교시 50분, 10분 휴식, 2교시 50분, 10분 휴식, 3교시 50분으로 진행 예정
- 강의 녹화영상은 업로드할 예정

과제

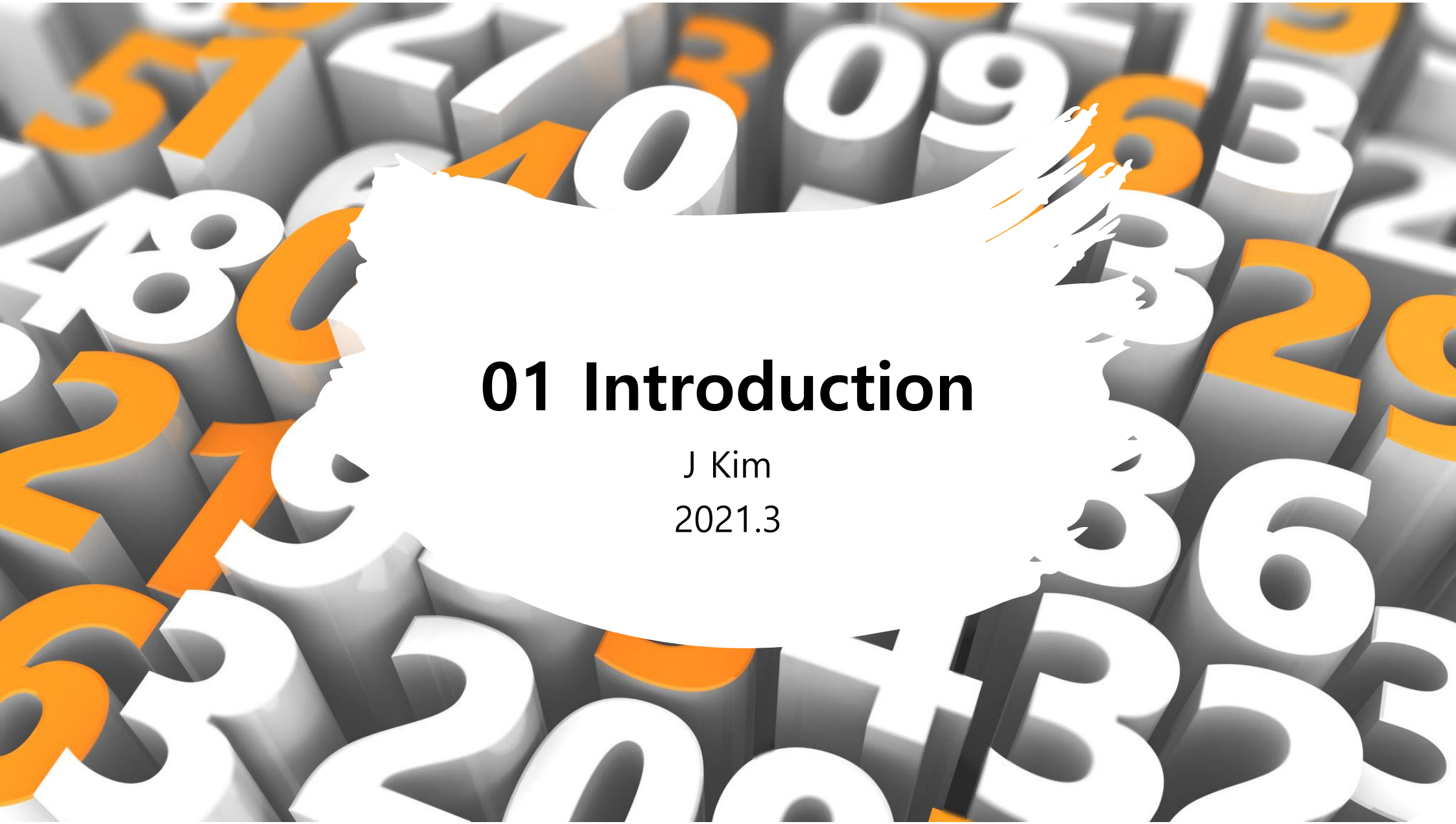
- 절대 남의 것을 베끼지 말고 혼자 힘으로!
- Canvas 지정한 곳에 pdf 파일로 변환하여 제출
- 마감시간이후에 제출된 과제는 감점
- 과제풀이는 동영상으로 업로드할 예정

평가

- 중간, 기말고사 평가시간은 60분내외
  - 중간고사: 4월26일(월) 오후 5:30~6:30
  - 기말고사: 6월14일(월) 오후 5:30~6:30
- 공학용 계산기 지참. 핸드폰에 내장된 계산기 사용 불허

Thank you!

Move on to 01 Introduction I



01 Introduction

J Kim

2021.3

Outline

통계적 기계학습
세 가지 자료

통계적 학습(Statistical learning)

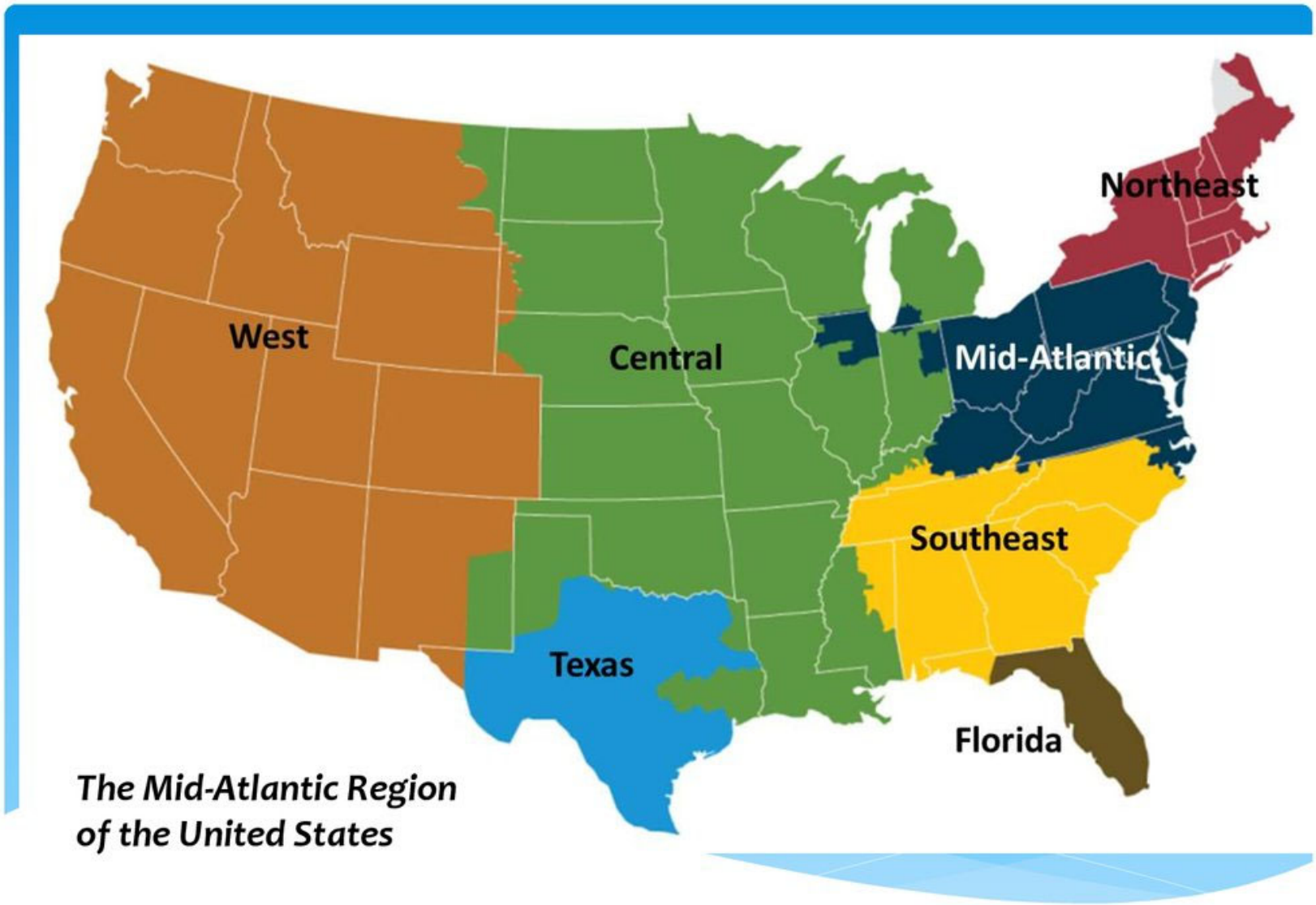
- 데이터를 이해하기 위한 방법들의 집합
- 지도(supervised)학습과 비지도(unsupervised)학습으로 구분
- 지도학습이란? 하나 혹은 둘 이상의 입력값(inputs)을 써서 결과값(output)을 예측하거나 추정하기 위해 통계 모형을 구축하는 것
- 비지도학습이란? 지도하는 결과값 없이 입력값만으로 입력 간의 관계나 구조를 마이닝(mining) 하는 것

임금(wage) 자료

- 중대서양(mid-atlantic) 지역에서 추출된 3,000명 남성 근로자들의 임금 자료
- 11개 변수로 구성된 data frame
- year: Year that wage information was recorded
- age: Age of worker
- education: A factor with levels 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad 5. Advanced
- wage: Workers raw wage

임금 자료

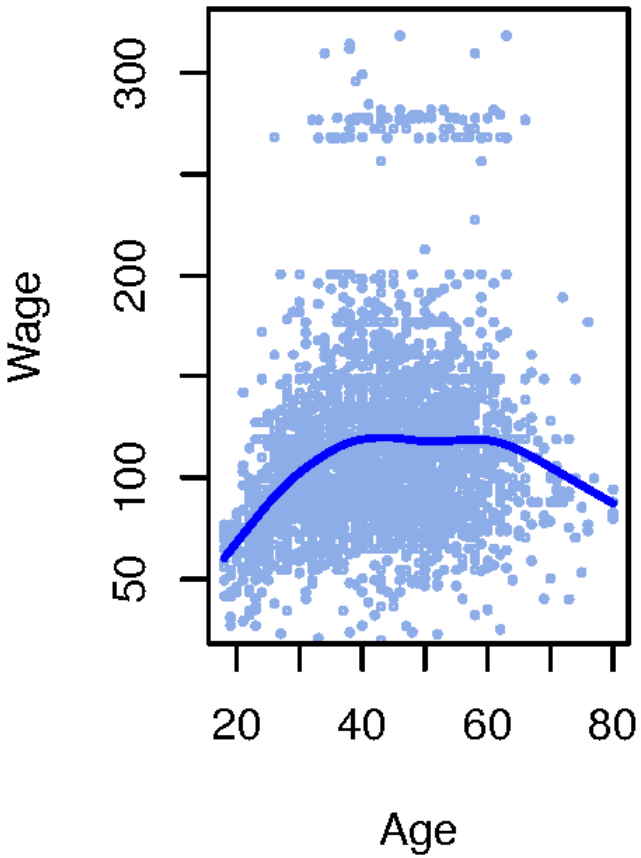
- maritl: A factor with levels 1. Never Married 2. Married 3. Widowed 4. Divorced 5. Separated
- race: A factor with levels 1. White 2. Black 3. Asian 4. Other
- region: Region of the country (mid-atlantic only)
- jobclass: A factor with levels 1. Industrial 2. Information
- health: A factor with levels 1. <=Good 2. >=Very Good
- health_ins: A factor with levels 1. Yes 2. No
- logwage: Log of workers wage

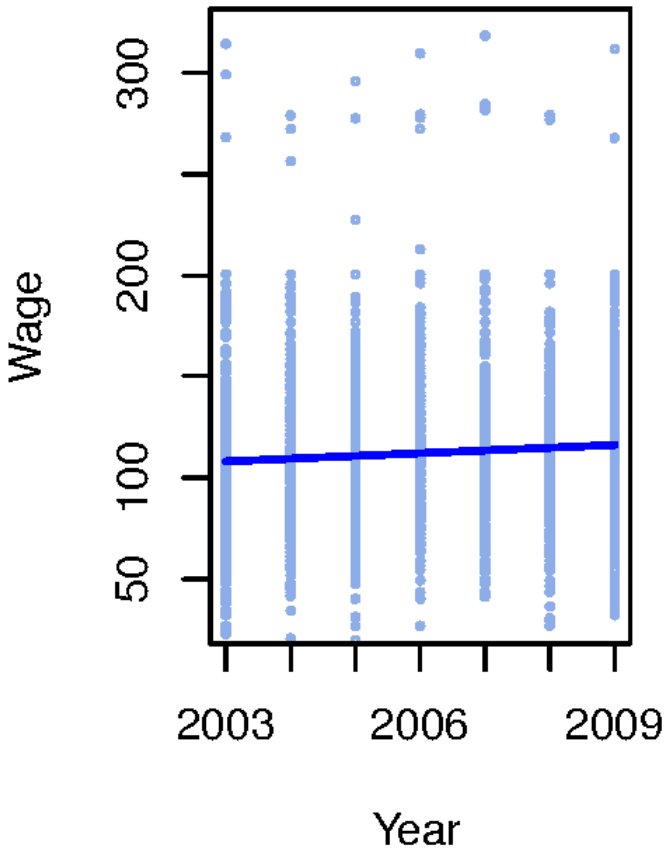


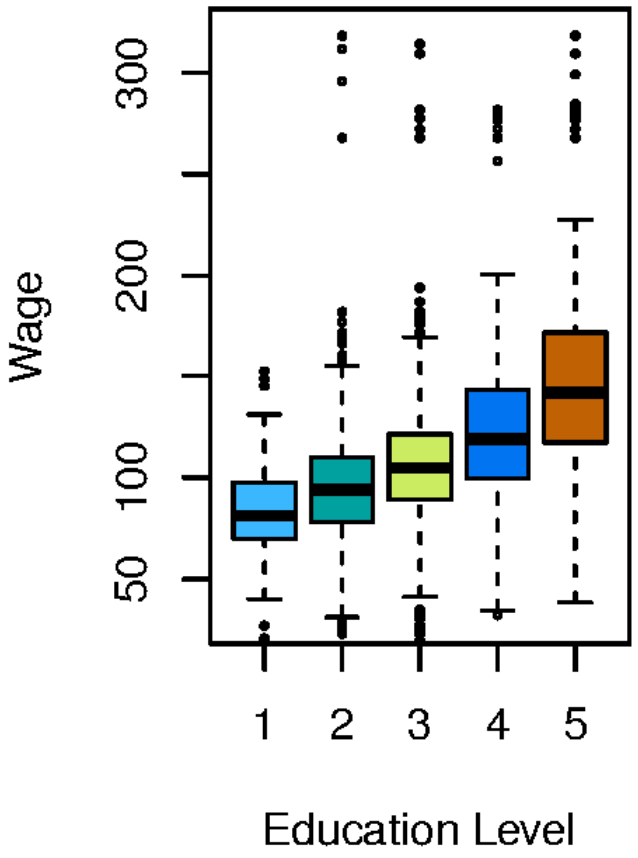
The Mid-Atlantic Region of the United States

임금 자료

- 연령, 교육수준, 관측연도가 임금과 어떻게 연관되어 있는가?
- 결과값: 연속적(continuous) 혹은 양적(quantitative)!
- 회귀(regression) 문제!


Wage
50
100
200
300
20
40
60
80
Age


Wage
50
100
200
300
2003
2006
2009
Year


Wage
50
100
200
300
1
2
3
4
5
Education Level

임금 자료: 결과 들여다보기

- 연령만으로 임금을 예측할 수 있을까? No. 유의미한 변동이 존재
- 2003년과 2009년 사이에 대략 만 불 증가했으며 선형적인 추세
- 교육수준이 높을수록 임금도 많음
- 3장에서 다룸!

증권시장(stock market) 자료

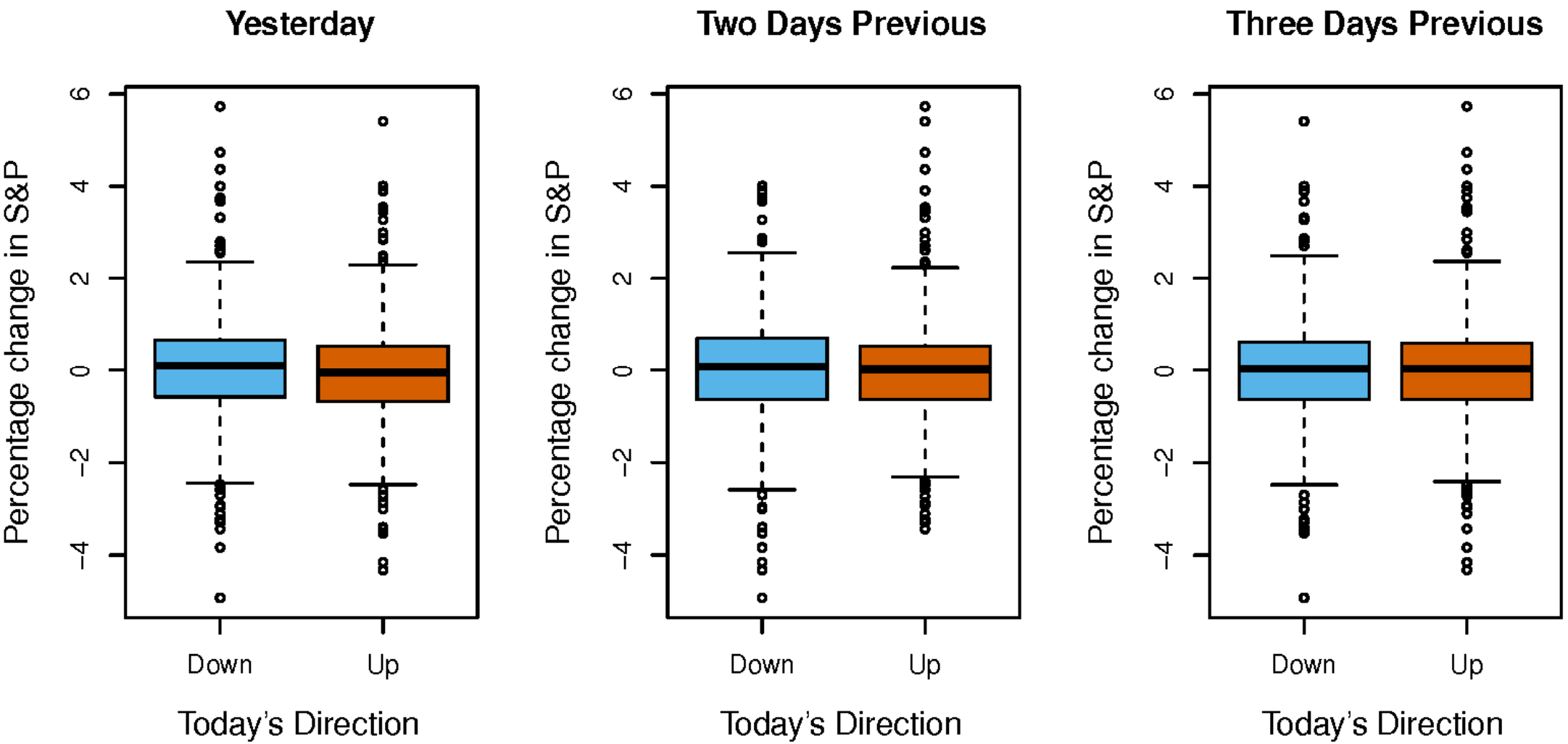
- 2001년부터 2005년까지 스탠더드 & 푸어스(Standard & Poor; S&P) 500 지수의 일별 수익률(%)
- 9개 변수, 1250개 관측으로 구성된 data frame

증권시장 자료

- Lag1: Percentage return for previous day
- Lag2: Percentage return for 2 days previous
- Lag3: Percentage return for 3 days previous
- Direction: A factor with levels Down and Up
- Year: The year that the observation was recorded
- Lag4: Percentage return for 4 days previous
- Lag5: Percentage return for 5 days previous
- Volume: Volume of shares(주식) traded
- Today: Percentage return for today

증권시장 자료

- 지난 5일간의 수익률을 바탕으로 오늘 수익률이 증가할 확률은 얼마인가?
- 결과값: 범주형(categorical) 혹은 질적(qualitative)!
- 분류(classification) 문제!

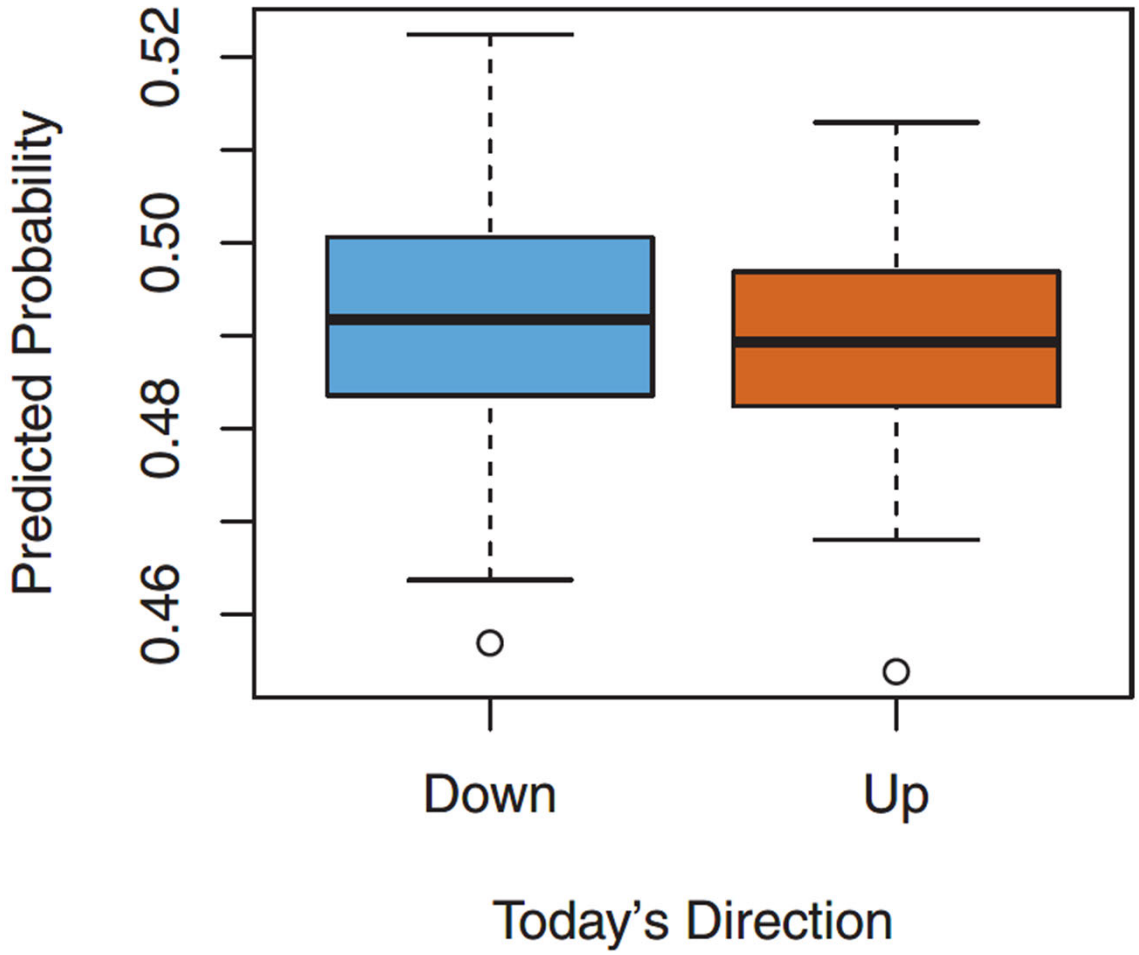

Yesterday
Two Days Previous
Three Days Previous
Percentage change in S&P
6
4
2
0
−2
−4
Down
Up
Today's Direction

증권시장 자료: 결과 들여다보기

- 하루 전날 수익으로 오늘의 수익을 예측하는 것은 어려움!
- 2, 3일 전과는 미세한 관계가 있음
- 4장에서 다룸!

증권시장 자료: 결과 들여다보기

- 이차판별분석(quadratic discriminant analysis) 결과 예시
- 2001-2004 자료만 사용해서, 2005년에 Down할 확률을 예측. 대략 60%정도 예측과 일치
- 4.6.4절 참조: Down 101일 중 30일을 Down으로 예측, Up 151일 중 121일을 Up으로 예측

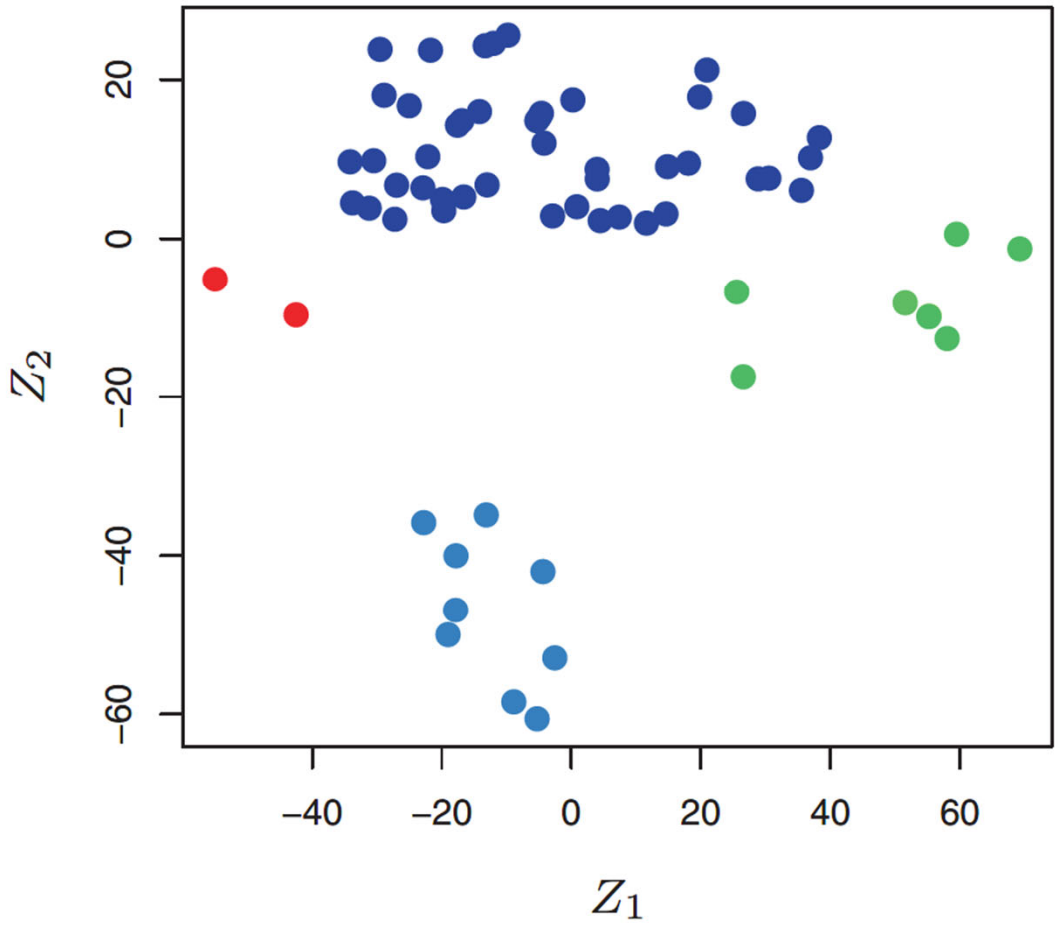

Predicted Probability
0.46
0.48
0.50
0.52
Down
Up
Today's Direction

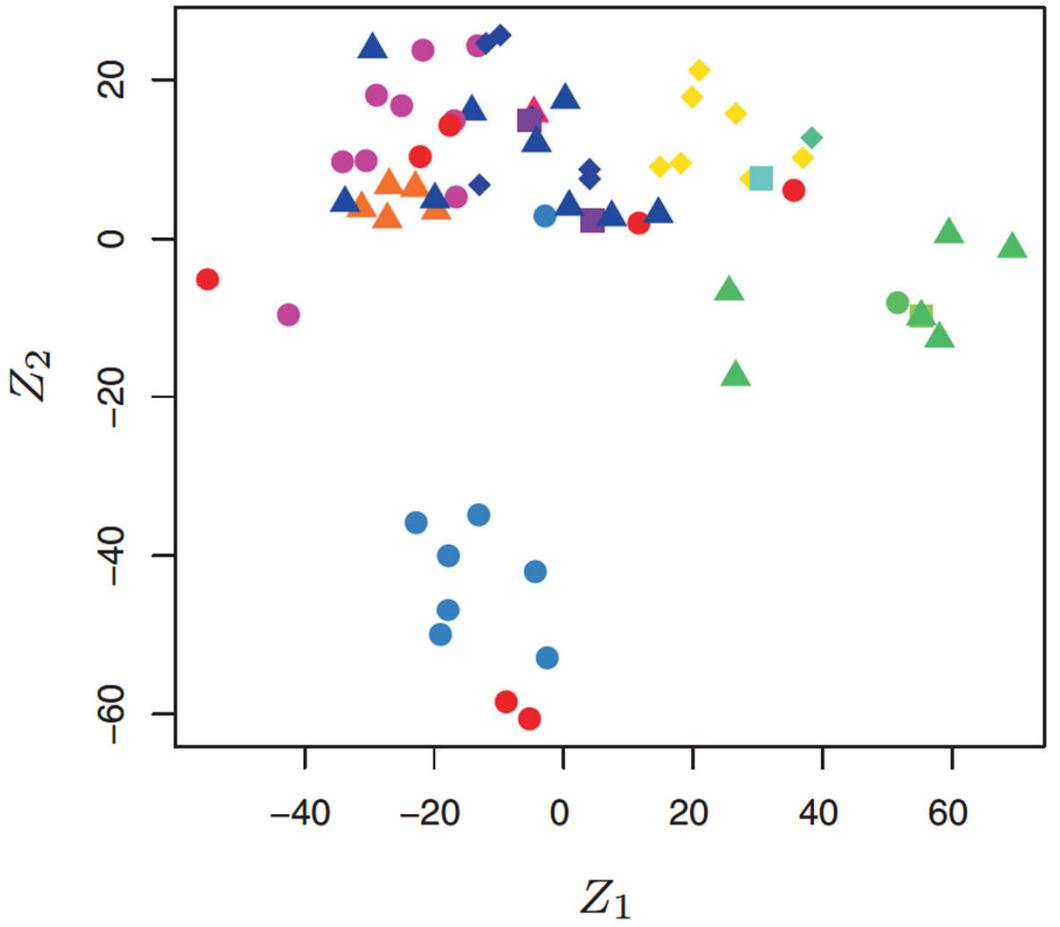
유전자 발현(gene expression) 자료

- 입력값만 관측!
- NCI60 자료
- 64개의 암세포(cancer cell lines)에 대한 6830 유전자의 발현값
- data와 labs으로 구성된 list
  - Data: a 64 by 6830 matrix of the expression values
  - Labs: a vector listing the cancer types for the 64 cell lines
- 유전자 발현값을 써서 암세포들을 유사한 세포 끼리 군집을 만들 수 있을까?

유전자 발현 자료

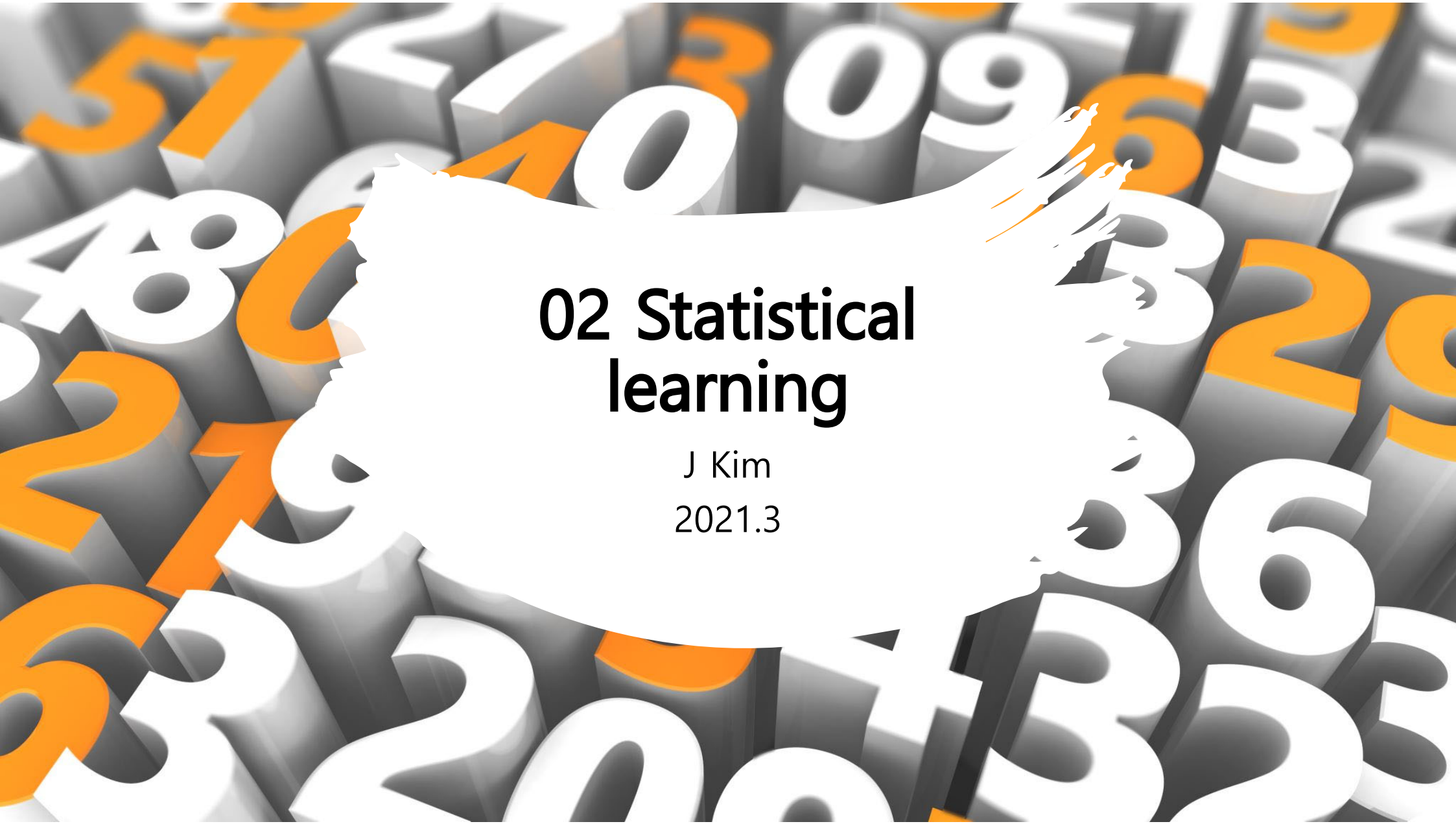
- 군집(clustering) 문제!
- 10장에서 다룸
- 입력변수가 많으면 시각화가 어려움!
- 차원 축소 필요!
  - 주성분(principal components) 2개만 고려
  - 군집(cluster)가 4개 이상 존재!
  - 실제 14개 종류의 암 존재
  - 암 정보를 전혀 사용하지 않았음에도 주성분분석(principal component analysis) 결과가 실제와 유사


Z_2
20
0
−20
−40
−60
−40
−20
0
20
40
60
Z_1


Z_2
20
0
−20
−40
−60
−40
−20
0
20
40
60
Z_1

Thank you!

Move on to 02 Statistical learning



02 Statistical learning

J Kim

2021.3

Outline

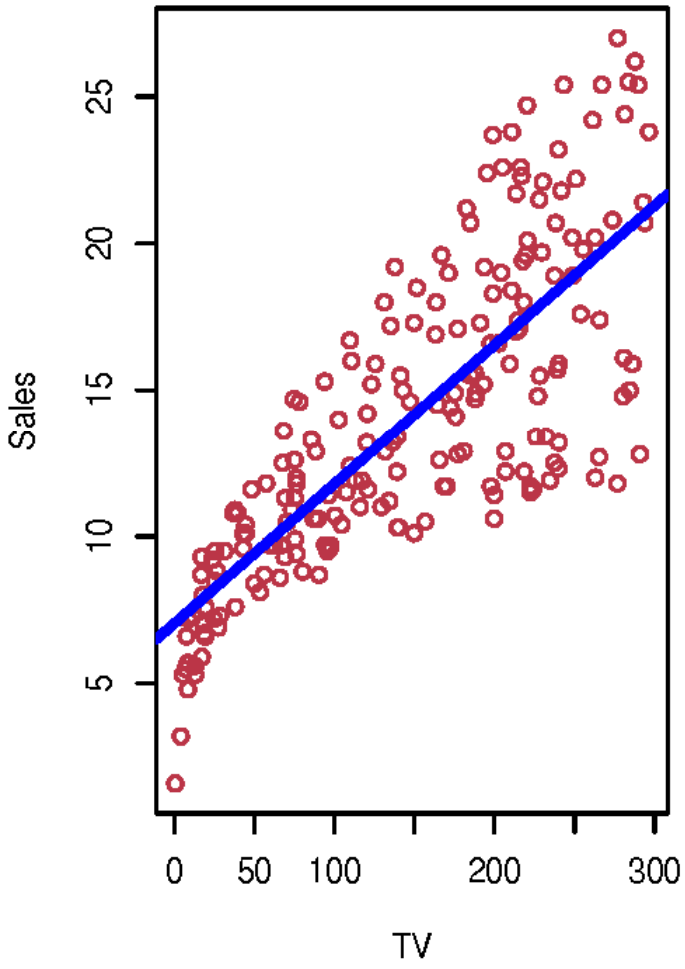
통계적 학습(statistical learning)이란?
모델의 정확성 평가: 평균제곱오차, 오류율

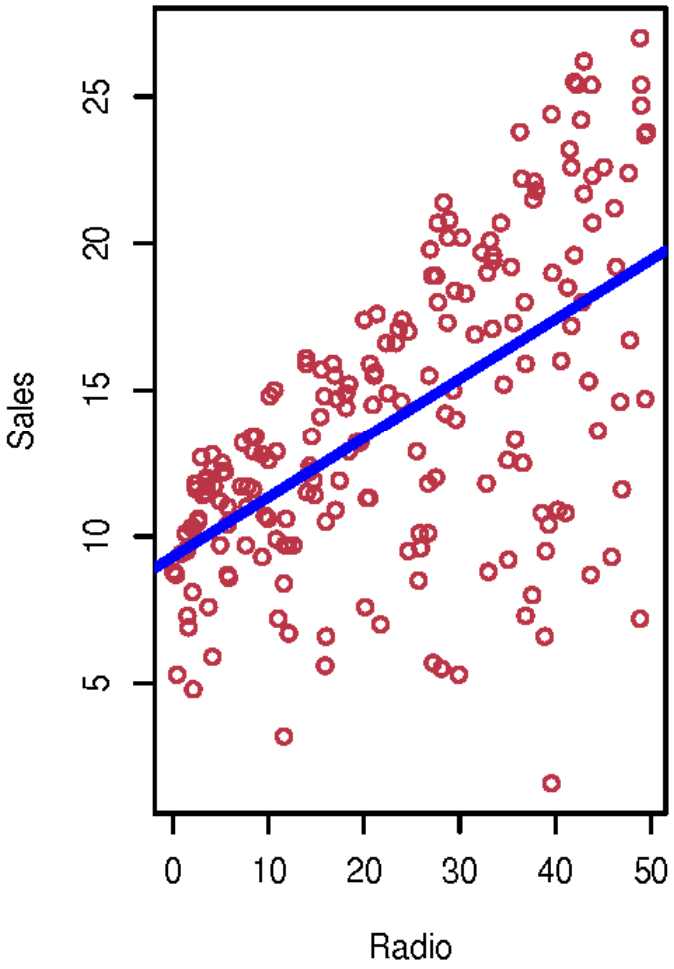
통계적 학습이란?

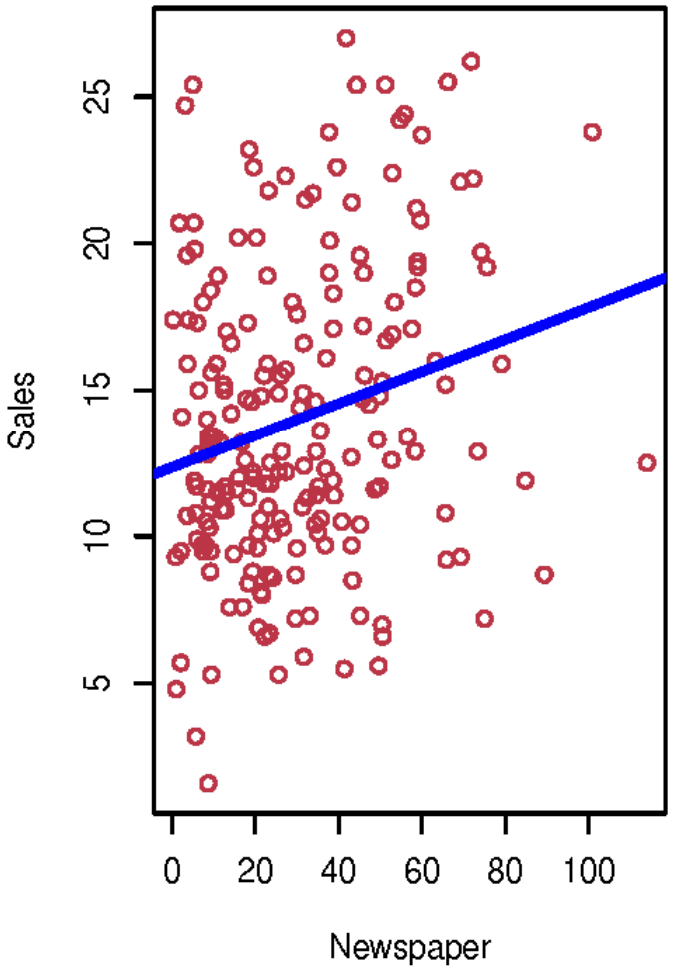
- $X_1, X_2, \ldots, X_p$: 입력변수(input variables), 예측변수(predictors), 독립변수(independent variables), 혹은 특성(features)
- Y: 결과변수(output variable), 반응변수(response variable) 혹은 종속변수(dependent variable)
- Y와 $X = (X_1, X_2, \ldots, X_p)'$의 관계를 표현하는 일반적인 꼴:

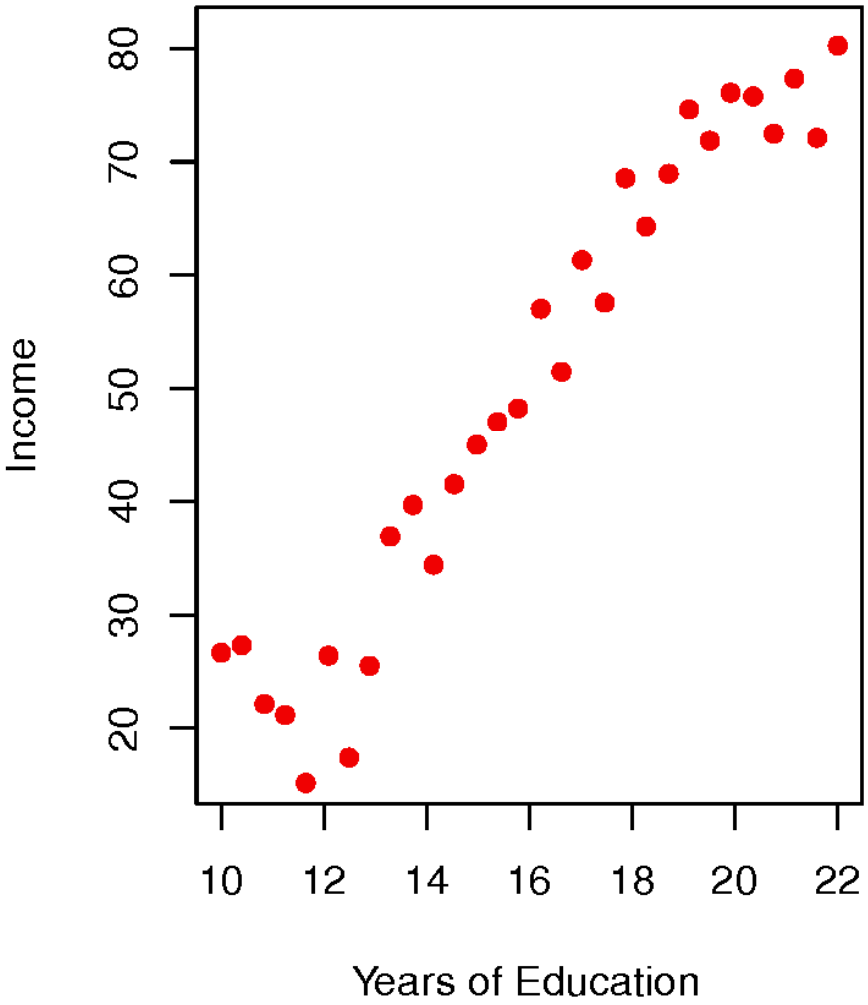
$$Y = f(X) + \epsilon$$

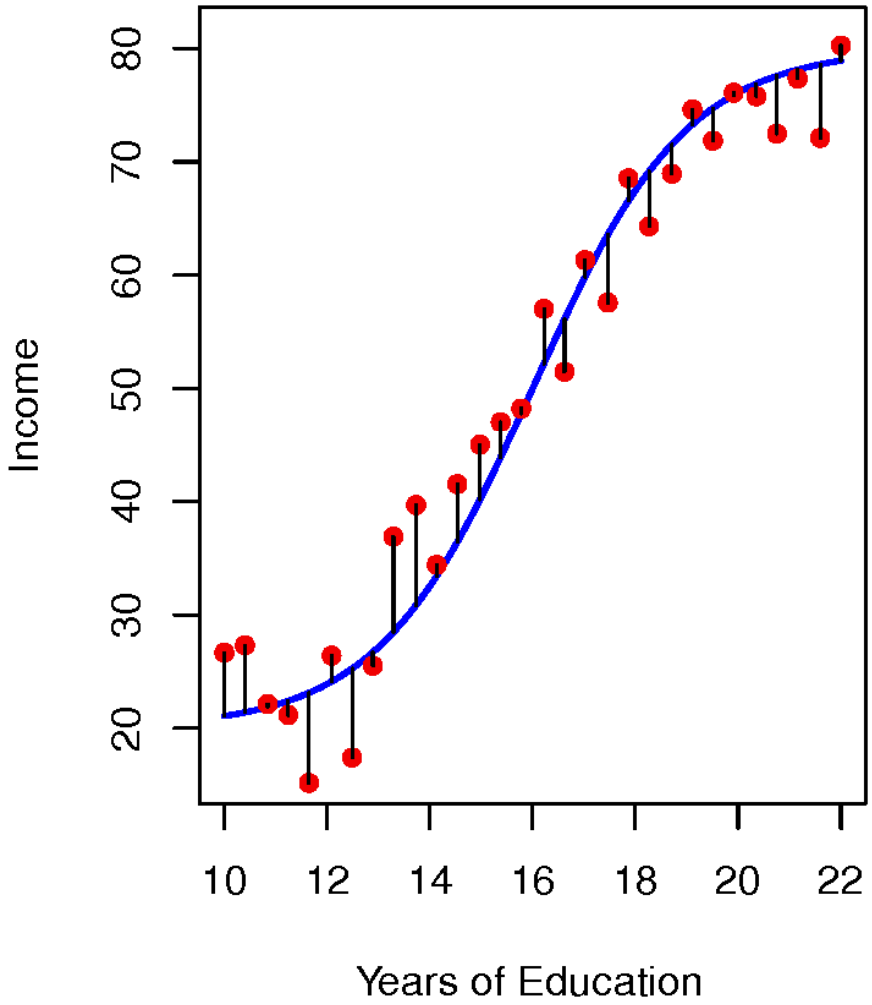
  - f: 미지의 함수. 체계적(systematic) 성분
  - ϵ: 랜덤 오차(random error) 평균이 0이고 예측변수와 독립
- f를 추정하기 위한 방법들의 모임!

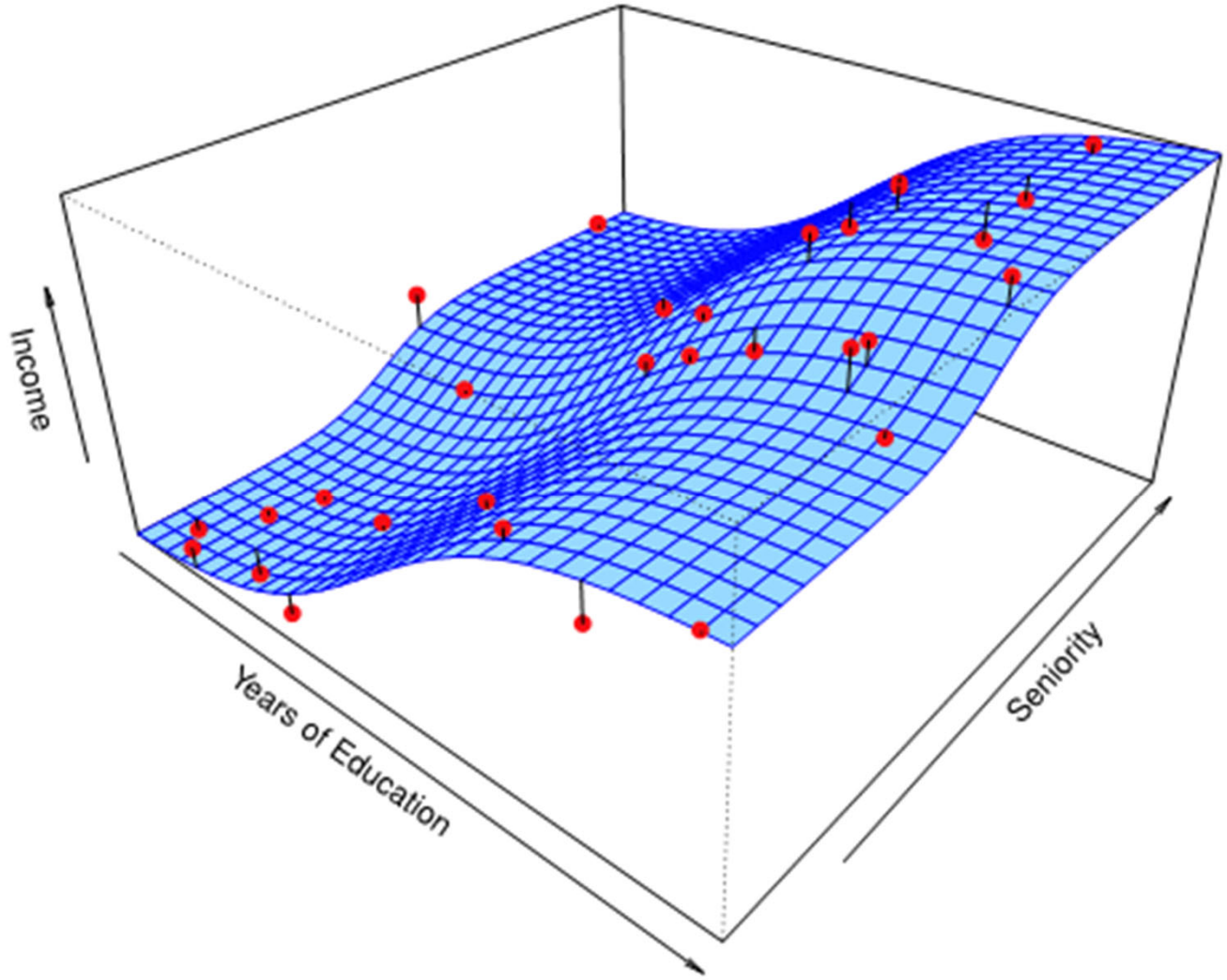

Sales
5
10
15
20
25
0
50
100
200
300
TV


Sales
5
10
15
20
25
0
10
20
30
40
50
Radio


Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper


Income
20
30
40
50
60
70
80
10
12
14
16
18
20
22
Years of Education


Income
20
30
40
50
60
70
80
10
12
14
16
18
20
22
Years of Education


Income
Years of Education
Seniority

왜 f를 추정하는가?

- 두 이유: 예측(prediction)과 추론(inference)
- 예측
  - 왜? X는 쉽게 얻을 수 있지만 Y는 그렇지 못함! (예: 약의 이상반응 여부)
  - 어떻게? $\hat{Y} = \hat{f}(X)$
  - $\hat{f}$은 검은 상자(black box)로 다룰 수 있음

$\hat{Y}$ 은 얼마나 정확한가?

- $\hat{Y}$의 정확도(accuracy)는 축소가능 오차(reducible error)와 축소불가능 오차(irreducible error)에 의존
- $\hat{f}$이 f에 대한 완전한 추정은 아님! 다만 더 적합한 통계적 학습 방법을 써서 정확도를 높일 수 있음
- $\hat{f} = f$라 할지라도 $\hat{Y} = f(X)$는 여전히 오차를 포함. Y는 ϵ의 함수이기도 하며 이는 X로 예측될 수 없어서! 즉 ϵ의 변동(variability)의 크기에도 의존

왜 축소불가능 오차는 존재하는가?

- 측정하지 못한 변수(unmeasured variables)가 있을 수도
- 측정 불가능한 변동(unmeasurable variation)이 있을 수도

평균제곱오차(mean squared error)

- Remarks: : $\hat{f}$과 $X = x$가 주어졌을 때,

$$E\left[\left(Y - \hat{Y}\right)^2 \middle| X = x\right] = \left[f(x) - \hat{f}(x)\right]^2 + Var(\epsilon)$$

  - 전자를 줄일 수 있는 방법을 학습
  - 후자가 정확도의 상한(upper bound)을 결정

왜 f를 추정하는가?

- 추론
  - 예측변수가 변함에 따라 반응변수가 어떻게 변하는지에 관심
  - 더 이상 $\hat{f}$은 검은 상자가 아님!
- 관심 질문
  - 반응변수와 연관된 예측변수는 무엇인기? 중요한 예측변수를 찾음
  - 반응변수와 개별 예측변수의 관계는 양수인가? 음수인가?
  - 반응변수와 개별 예측변수의 관계는 선형적인가? 좀 더 복잡한가?

예측 and/or 추론

- 문제에 따라 예측만 관심, 추론만 관심, 혹은 둘 다 관심이 있을 수도!
- 인구학적 변수의 값으로 우편 홍보에 긍정적으로 행동할 소비자를 찾아내고자 하는 자료 ⇒ 예측에 관심
- 광고 자료
  - 어떤 매체가 매출에 기여하는가?
  - 어떤 매체가 매출에 가장 큰 기여를 하는가?
  - TV 광고를 늘리면 매출이 얼마나 늘어나는가?

  ⇒ 추론에 관심

전략

- 추론에 초점을 맞추면 f는 단순하고 해석이 가능한(interpretable) 모형으로! ⇒ 예: 선형모형
- 예측에 초점을 맞추면 f는 해석은 어렵지만 높은 정확도를 가진 복잡한 모형으로 ⇒ 예: 비선형모형

어떻게 f를 추정하는가?

- 훈련세트(training set): $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
  - $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})' : x_{ij}$ 는 i-번째 개체의 j −번째 예측변수의 관측값
  - y_i: i −번째 개체의 반응변수의 관측값
- 목표: (X, Y)의 어떤 관측값에도 $Y \approx \hat{f}(X)$이 되도록 하는 $\hat{f}$ 를 찾음
- 두 방법: 모수적(parametric), 비모수적(non-parametric)

모수적방법

- 함수 꼴을 가정
  - 예: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ (선형) $\Rightarrow \beta_0, \beta_1, \ldots, \beta_p$ 만 추정하면 되며 간단한 편임
- f를 추정하는 문제가 모수를 추정하는 문제로 축소
- 예: 그림 2.4 (선형 모형)

비모수적 방법

- f의 꼴을 가정하지 않고 훈련세트에 가까운 f의 추정량을 찾음
- f의 추정 문제가 모수의 추정 문제로 축소 되지 않기 때문에 훈련세트의 크기가 커야!
- 예: 그림2.5 (부드러운 스플라인 모형), 그림2.6 (거친 스플라인 모형, 과적합)

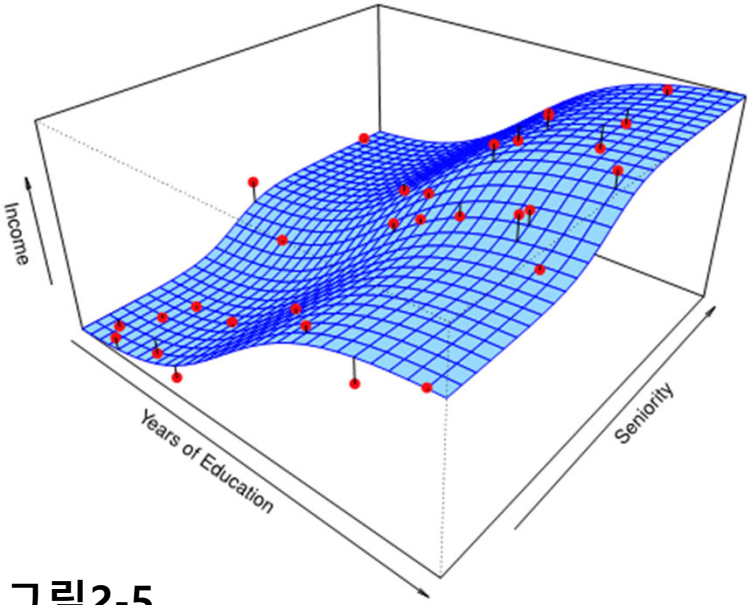


그림2-4

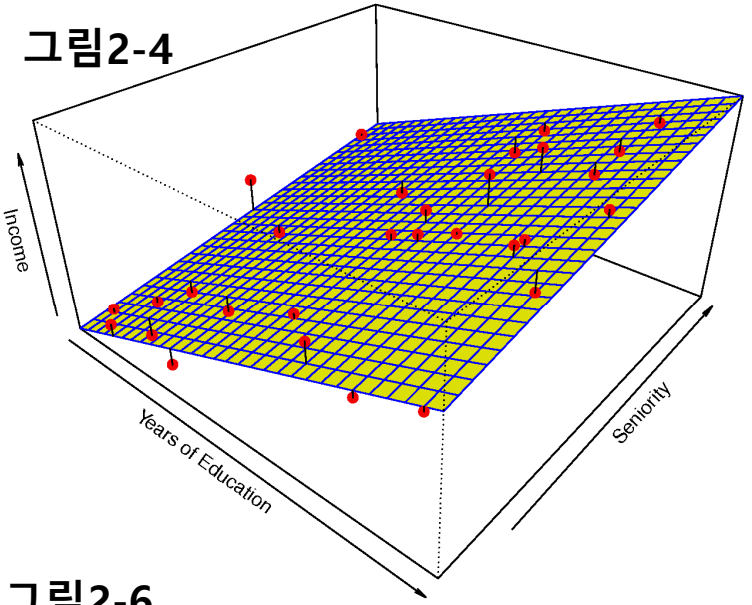


그림2-5

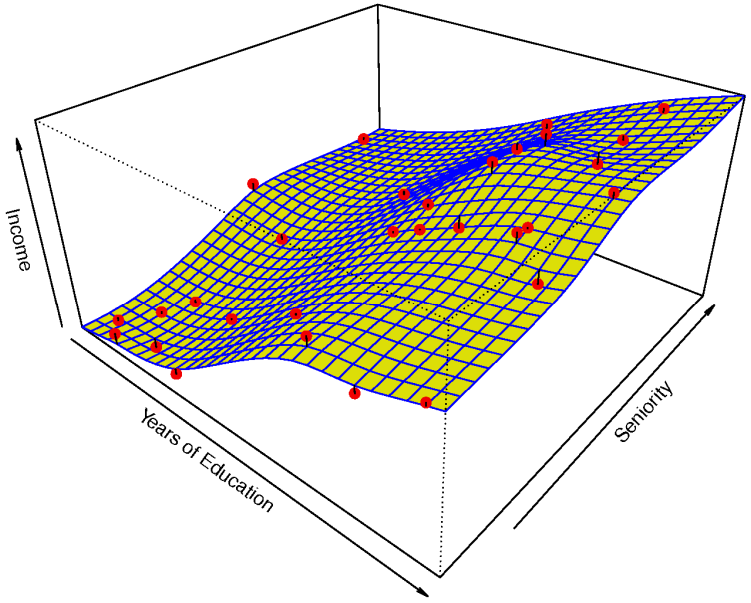
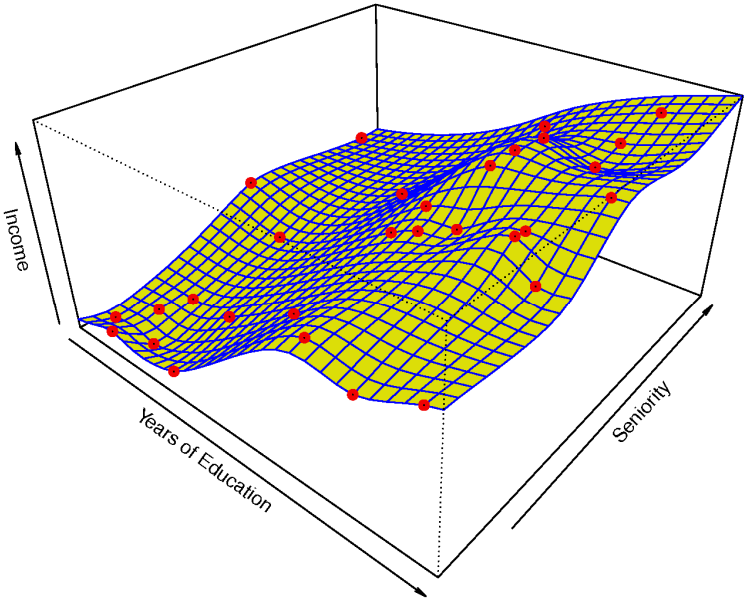
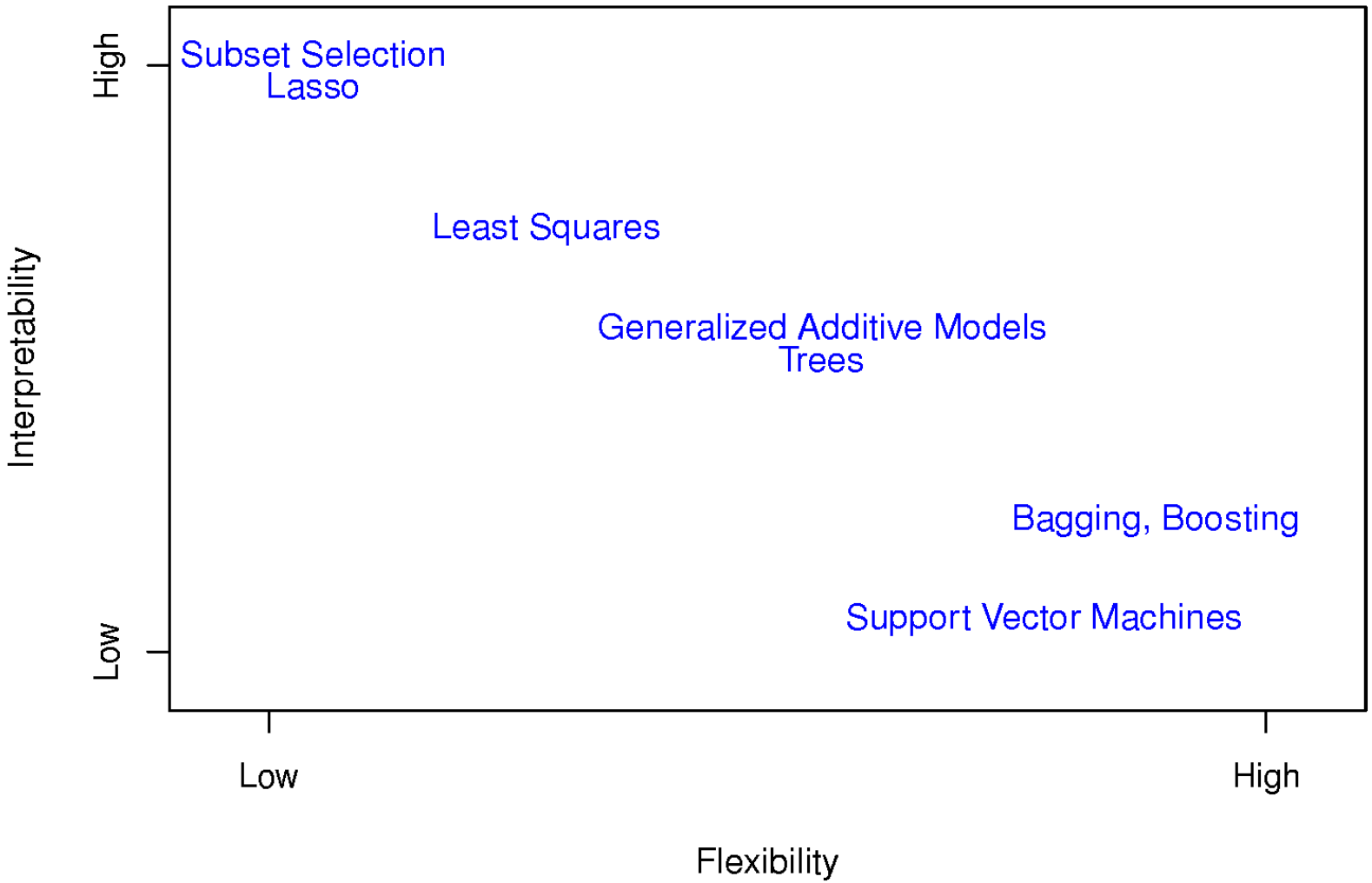


그림2-6



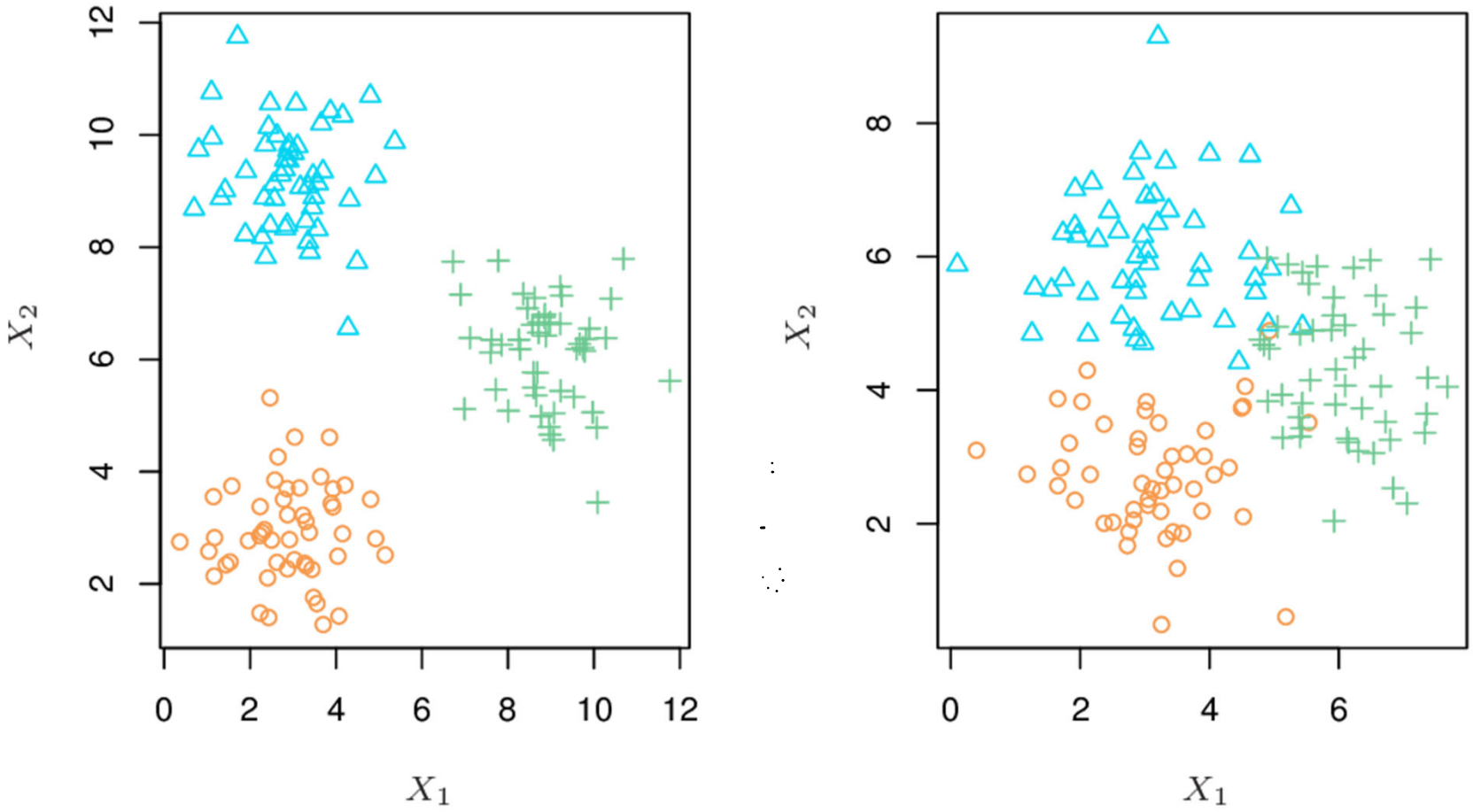
예측정확도(prediction accuracy)와 해석력(interpretability)의 교환

- 선형 모형-제한적(직선이나 평면으로 추정) vs. 스플라인 모형-유연(다양한 형태(shape)로 추정 가능)
- 왜 제한적 모형을 사용하는가? 추론에 관심을 둠: 해석력이 뛰어남 (예측력은 떨어짐)
- 해석적(제한적): 변수 선택, LASSO ≫ 최소제곱법 ≫ GAMs(일반화가법모형) ≫ 배깅, 부스팅, SVM(서포트 벡터 머신)


Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
High
Low
Interpretability
Low
High
Flexibility

지도학습(supervised learning)과 비지도학습(unsupervised learning)

- 지도학습: 훈련세트가 x_i에 연관된 y_i로!
  - 선형모형, 로지스틱 회귀모형, GAMs, 부스팅, SVM 등
- 비지도학습: 훈련세트가 x_i로만!
  - 변수들 혹은 개체들 간의 관계를 찾거나 집단으로 세분하는 데 목표를 둠
  - 주성분분석, 군집분석 등


X_2
X_1
0
2
4
6
8
10
12

X_2
X_1
0
2
4
6
8

회귀(regression)와 분류(classification)

- 반응변수: 양적, 질적(범주형)
- 예
  - 회귀만: 선형모형
  - 분류만: 로지스틱 회귀모형
  - 회귀 & 분류: KNN, 부스팅 등

적합 품질(quality of fit) 측정

- 왜 많은 통계적 학습 방법을 배워야 하는가? 모든 데이터에 대해 항상 우수한 단 하나의 모형은 없어서!
- 주어진 자료에 가장 좋은 결과를 주는 방법을 선택해야
- 아이디어: 반응변수의 예측값과 참값을 비교

적합 품질 측정: 회귀

- 훈련 MSE
  - 평균제곱오차(mean squared error, MSE)=$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{f}(x_i)\right)^2$
- 테스트 MSE: 테스트 세트에 속한 개체 (x_0, y_0)에 대해 $\left(y_0-\hat{f}(x_0)\right)^2$를 구한 후 평균 계산
- 테스트 MSE를 작게 하는 모형을 선택

적합 품질 측정

- 테스트 세트가 없으면 훈련 MSE로 비교. 다만 훈련 MSE를 가장 작게 하는 학습 방법이 항상 테스트 MSE를 가장 작게 하지 않을 수도!
- 모형의 유연성이 늘어날수록 (자유도가 높을수록) 훈련 MSE는 작아짐. 그러나 테스트 MSE는 항상 그렇지 않음 ⇒ 과적합!
- 훈련 MSE가 테스트 MSE보다 작음

적합 품질 측정

- 예: 가상(simulated) 자료(그림2.9)
  - 테스트 MSE는 파란색 모형이 가장 작음!
  - 수평 점선: 축소불가능한 오차 $Var(\epsilon) = 1$이므로 테스트 MSE는 1보다 항상 큼. 파란색 모형이 최적(optimal) 모형에 가까움!
  - 녹색 모형은 과적합(overfitting)
  - "a least flexible model would have yielded a smaller test MSE"
- 두 예: 그림2.10(선형), 그림2.11(비선형)
- Remarks: 실제는 교차검증(cross-validation) 방법으로 테스트 MSE를 추정

그림2-9 가상자료. 왼쪽(검은색: True f(X), 오렌지색: 선형회귀모형, 파란색, 녹색: 스플라인), 오른쪽(녹색: 훈련 MSE, 빨간색: 테스트 MSE, 점선: 오차 분산=1)

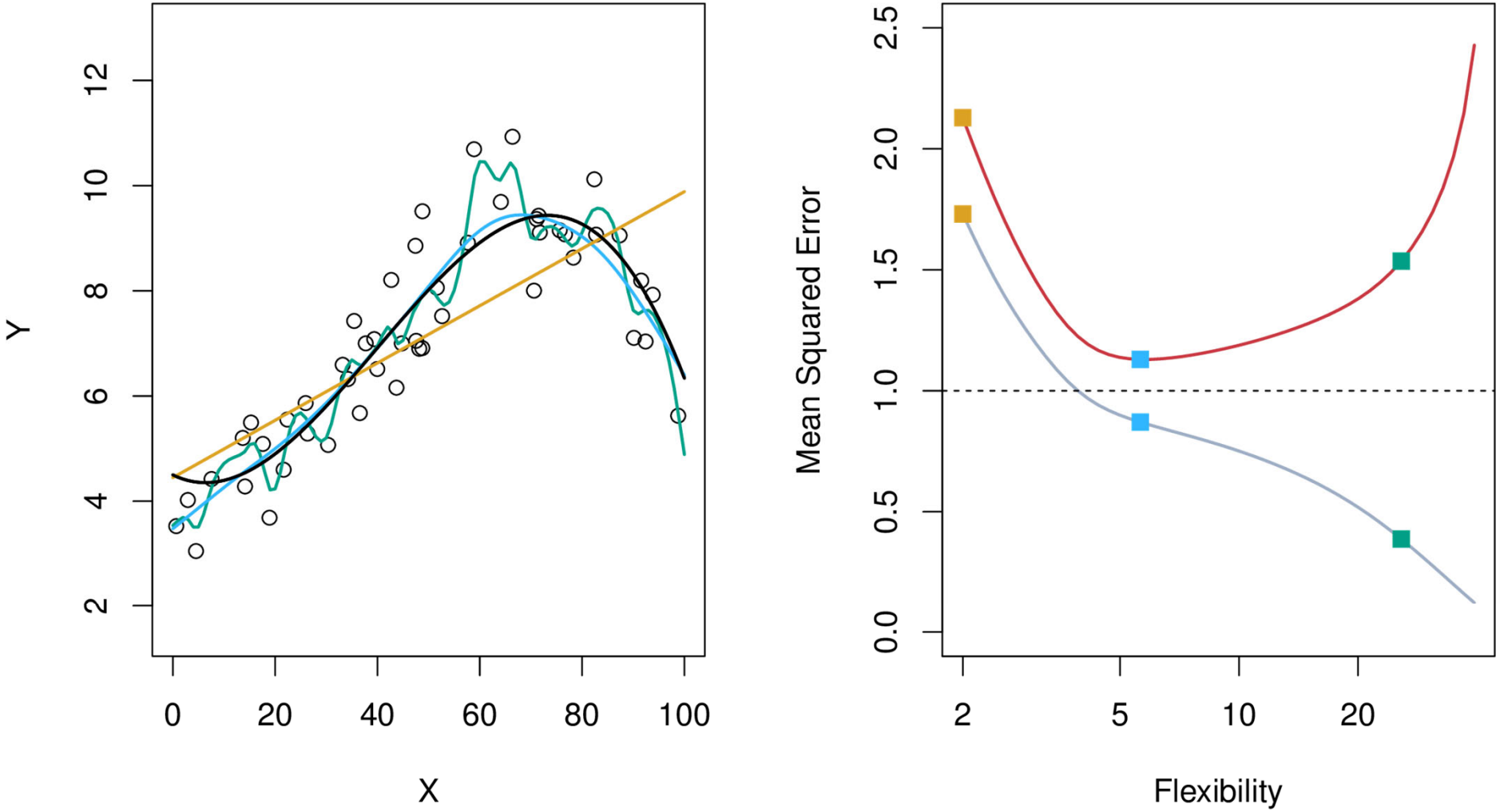


그림2-10 가상자료(선형). 왼쪽(검은색: True f(X), 오렌지색: 선형회귀모형, 파란색, 녹색: 스플라인), 오른쪽(녹색: 훈련 MSE, 빨간색: 테스트 MSE, 점선: 오차 분산=1)

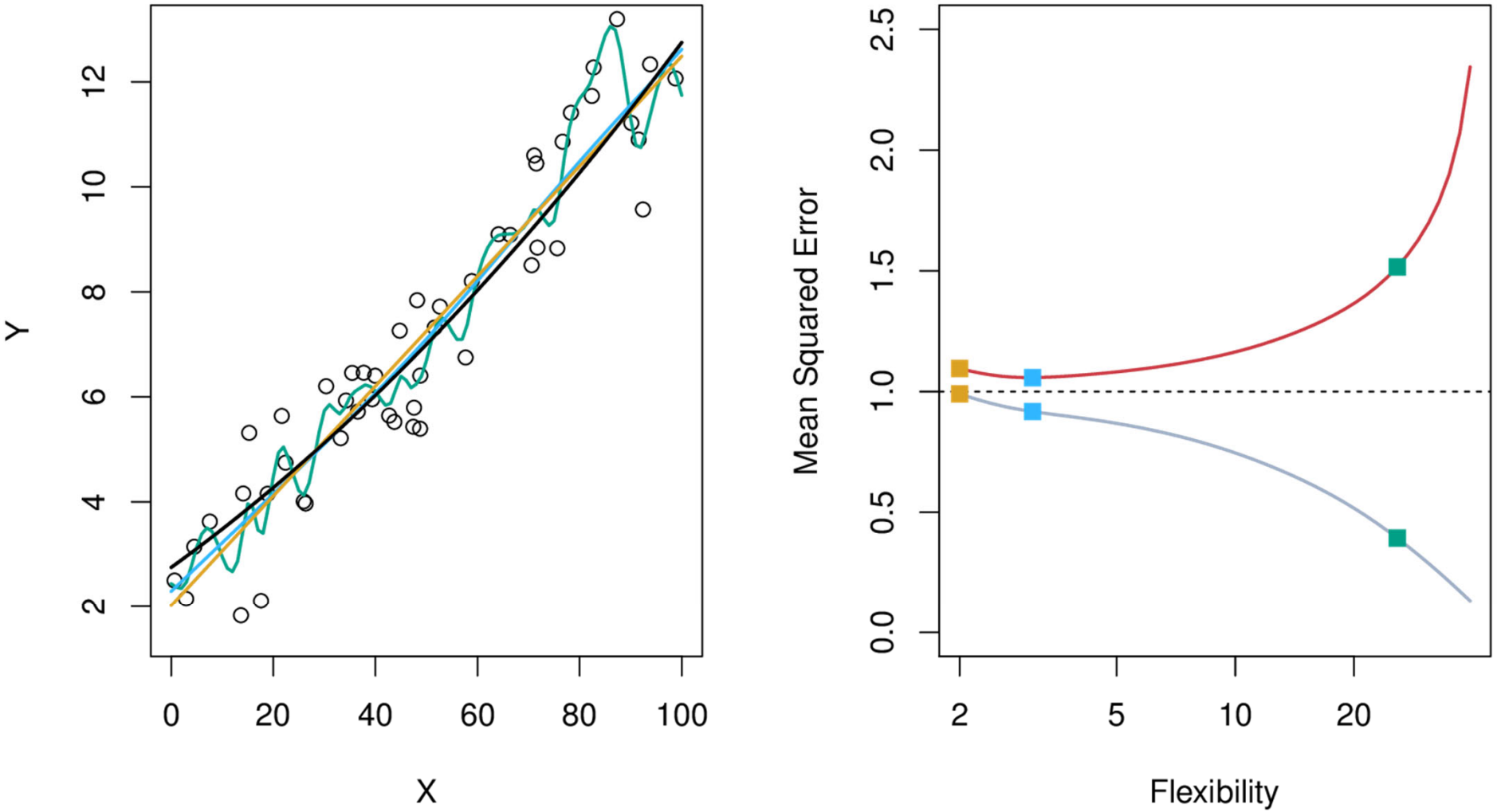
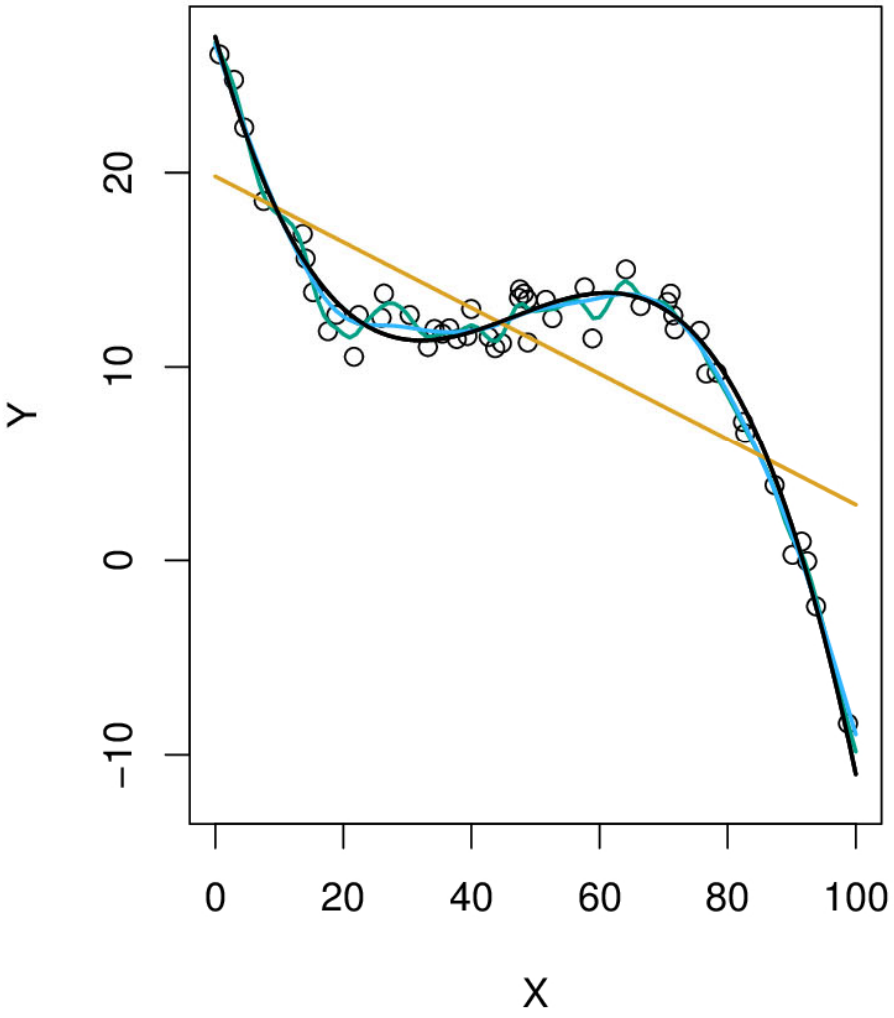
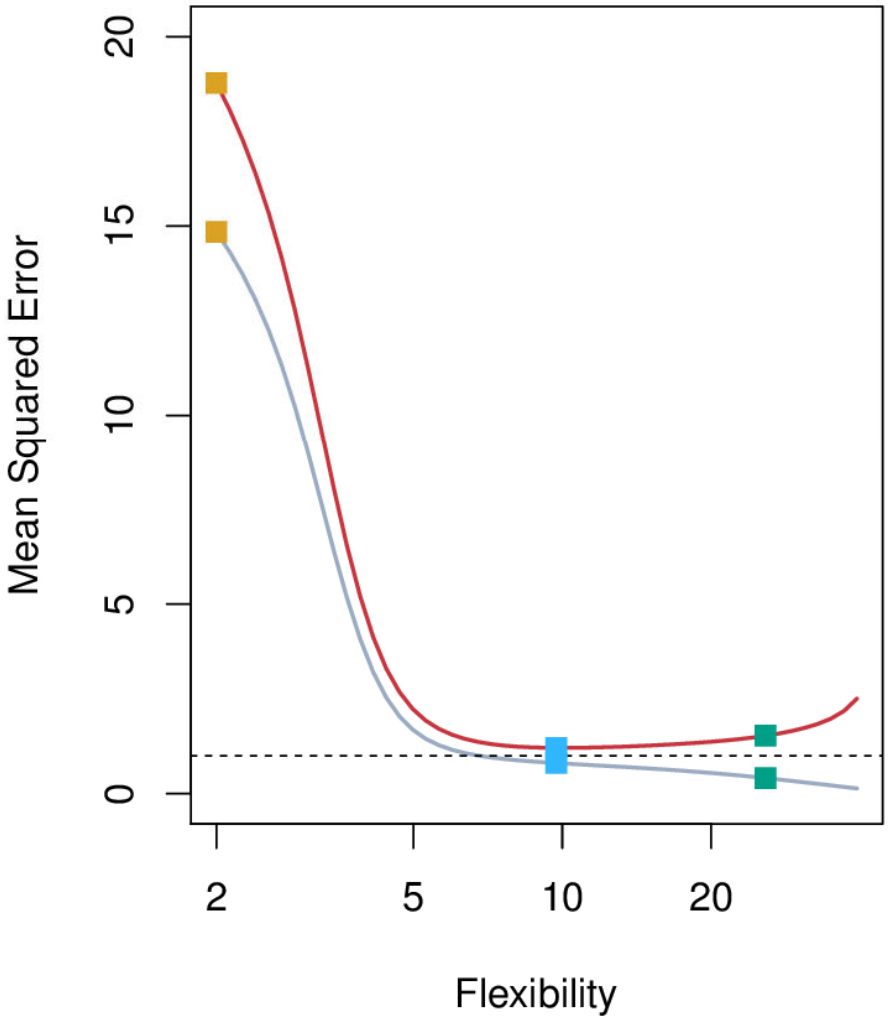


그림2-11 가상자료(비선형). 왼쪽(검은색: True f(X), 오렌지색: 선형회귀모형, 파란색, 녹색: 스플라인), 오른쪽(녹색: 훈련 MSE, 빨간색: 테스트 MSE, 점선: 오차 분산=1)





편의(bias)와 분산(variance)의 교환

- 기대(expected) 테스트 MSE: 주어진 x_0에 대해,

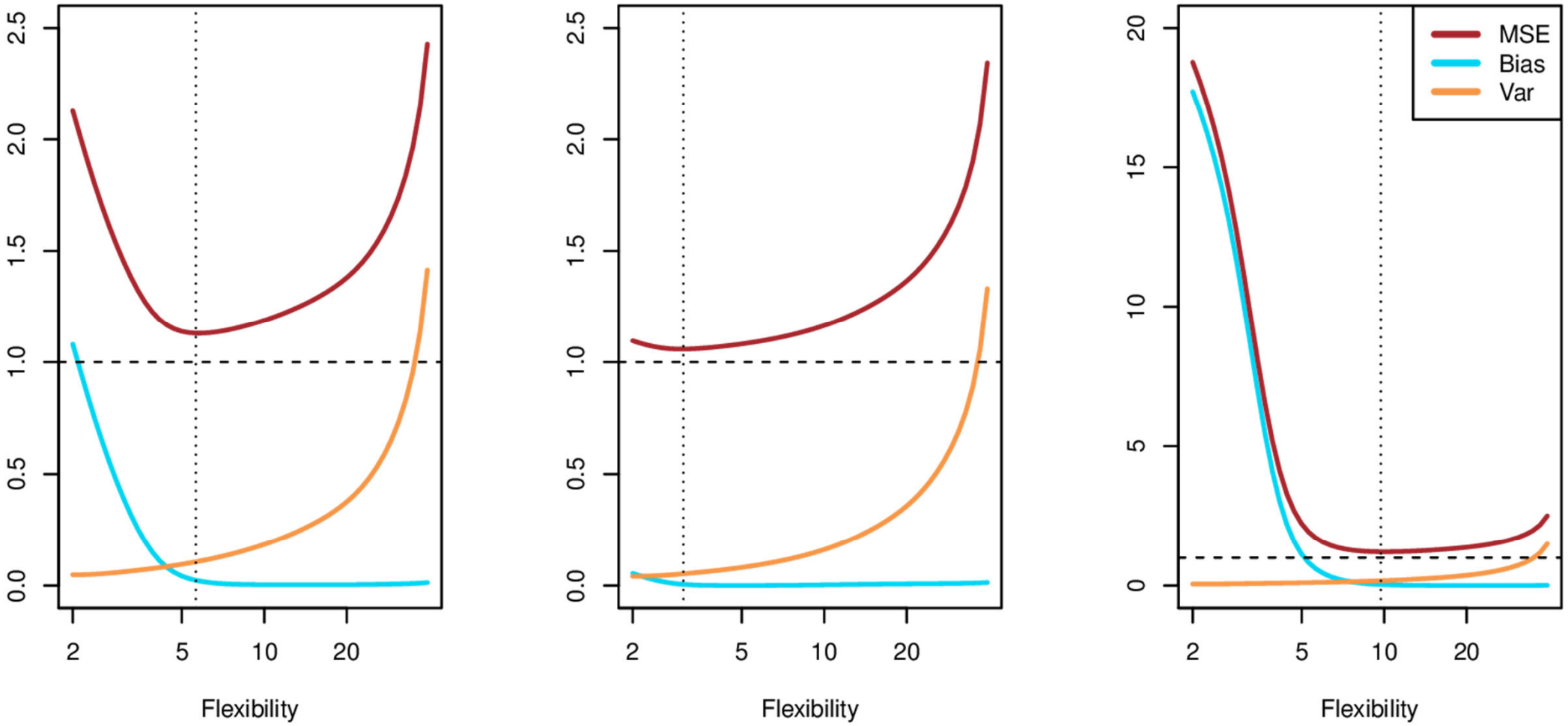
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left\{Bias\left(\hat{f}(x_0)\right)\right\}^2 + Var(\epsilon)$$

- 의미: 많은 훈련세트로 반복해서 f를 추정했을 때 x_0에서 얻어지는 MSE의 평균
- 통합(overall) 기대 테스트 MSE: 테스트 세트에 속한 개체 (x_0, y_0)에 대해$E\left(y_0 - \hat{f}(x_0)\right)^2$를 구한 후 평균 계산

편의와 분산의 교환

- 통계적 학습의 분산과 편의는 무슨 뜻인가?
  - 분산은 다른 훈련세트로 f를 추정했을 때 $\hat{f}$의 변동 ⇒ 작을수록 이상적. 유연성이 높을수록 분산은 커짐
  - 편의는 추정오차에 해당 ⇒ 작을수록 이상적. 유연성이 낮을수록 편의는 커짐
- 테스트 MSE의 증감은 분산과 편의의 상대적 변화 속도에 의존

그림2-12 가상 자료(그림2.9-2.11). 수평 점선(오차 분산=1), 수직 점선(테스트 MSE가 가장 작은 유연성 정도)



분류 문제

- MSE 대신 오류율을 계산: 오류율(error rate)=$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$
  - $\hat{y}_i$: $\hat{f}$으로부터 예측된 범주(라벨)
- 오분류된 비율(fraction of incorrect classification)
- 테스트 오류율: 테스트 세트에 속한 개체 (x_0, y_0)에 대해 $I(y_0 \neq \hat{y}_0)$를 구한 후 평균 계산
- 테스트 오류율을 작게 하는 모형을 선택

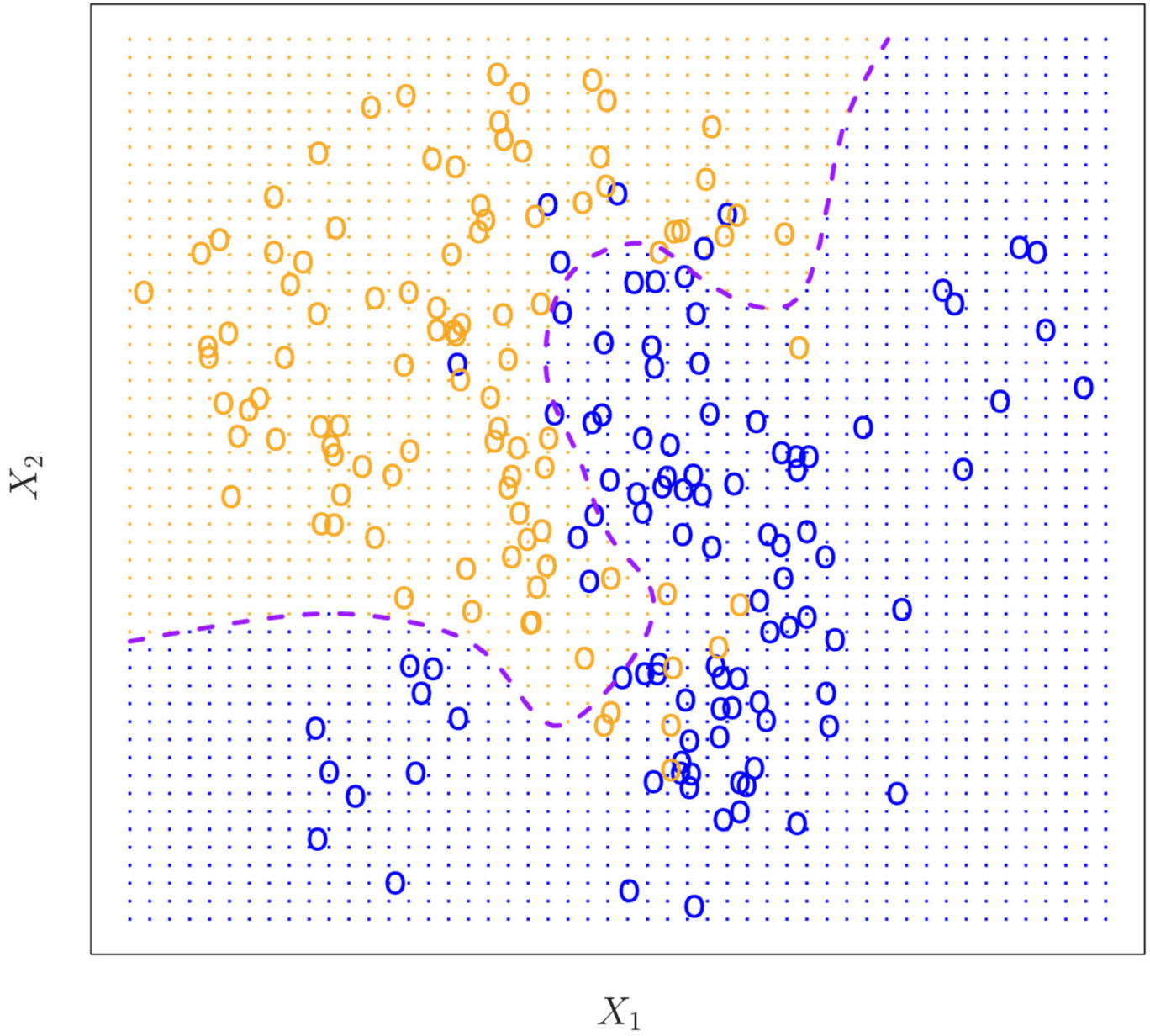
베이즈 분류기(Bayes classifier)

- 아이디어: 주어진 예측변수 값에 대해 가장 그럴듯한 범주(class)로 분류
- 예측변수의 값 x_0에 대해 j-번째 범주에 속할 조건부 확률 즉, $P(Y = j|X = x_0)$을 가장 크게 하는 범주로 분류

베이즈 분류기: 2-범주(범주: 1,2) 문제

- 만약 $P(Y = 1|X = x_0) > 0.5$이면 범주 1로 분류; 그렇지 않으면 범주 2로 분류
- 베이즈 결정 경계(Bayes decision boundary): $\{x|P(Y = 1|X = x) = 0.5\}$ ⇒ 보라색 점선

그림2-13 가상 자료. 각 범주에서 100개 자료. K=3. 보라색: 베이즈 결정 경계



베이즈 분류기

- Remarks 1:
  - $X = x_0$에 대한 베이즈 오류율 $= 1 - \max_j P(Y = j|X = x_0)$
  - 통합(overall) 베이즈 오류율 $= 1 - E\left(\max_j P(Y = j|X)\right)$: "expectation"은 모든 가능한 X의 값에 대한 평균"을 의미
- Remarks 2: 가상 자료에서 통합 베이즈 오류율은 $= 0.1304$
  - 왜 0보다 큰가? 두 모집단이 서로 겹치는 부분이 있음
  - 회귀 문제에서 축소불가능한 오류와 유사

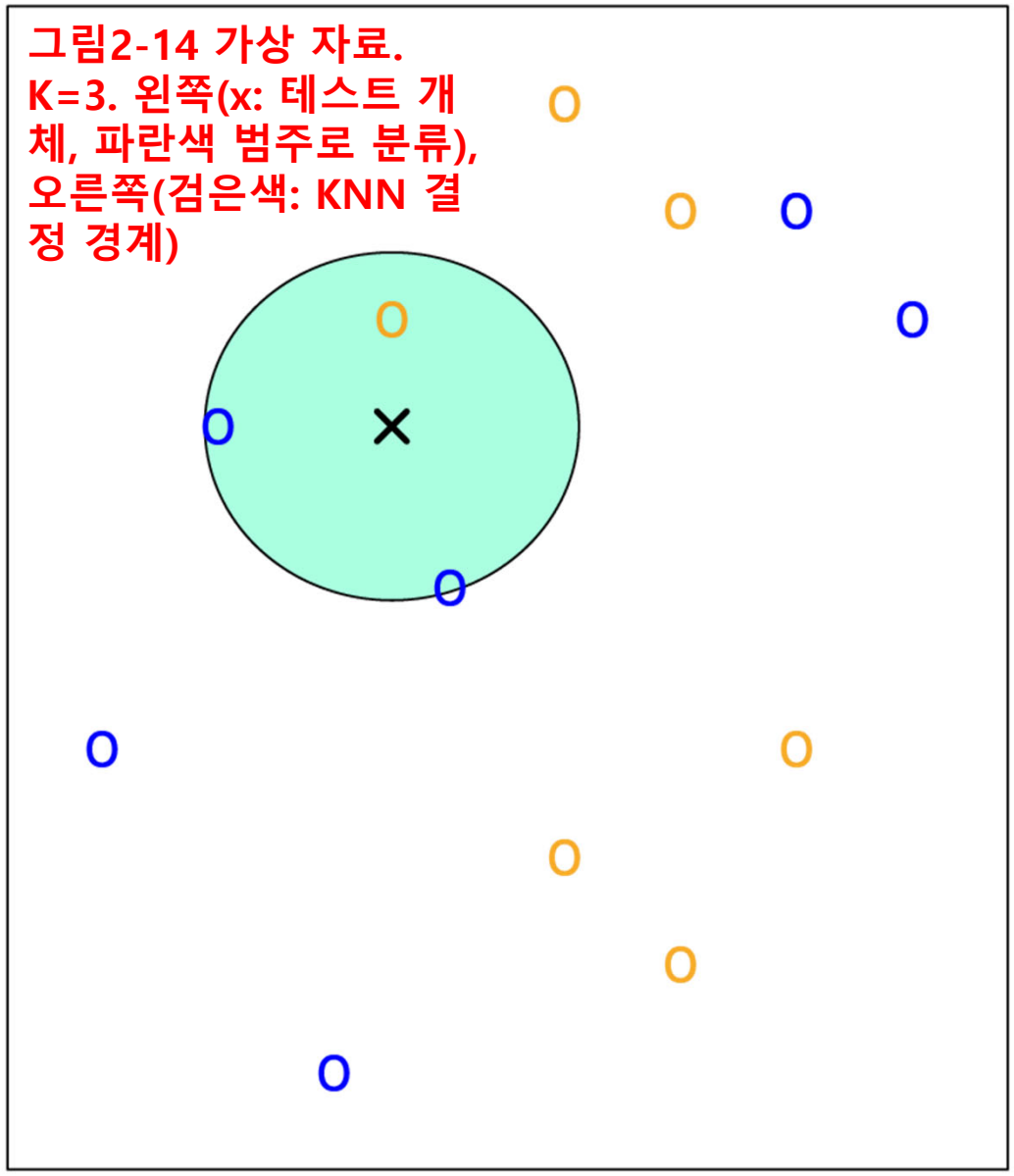
K-최근접 이웃 분류기(K-nearest neighbor classifier)

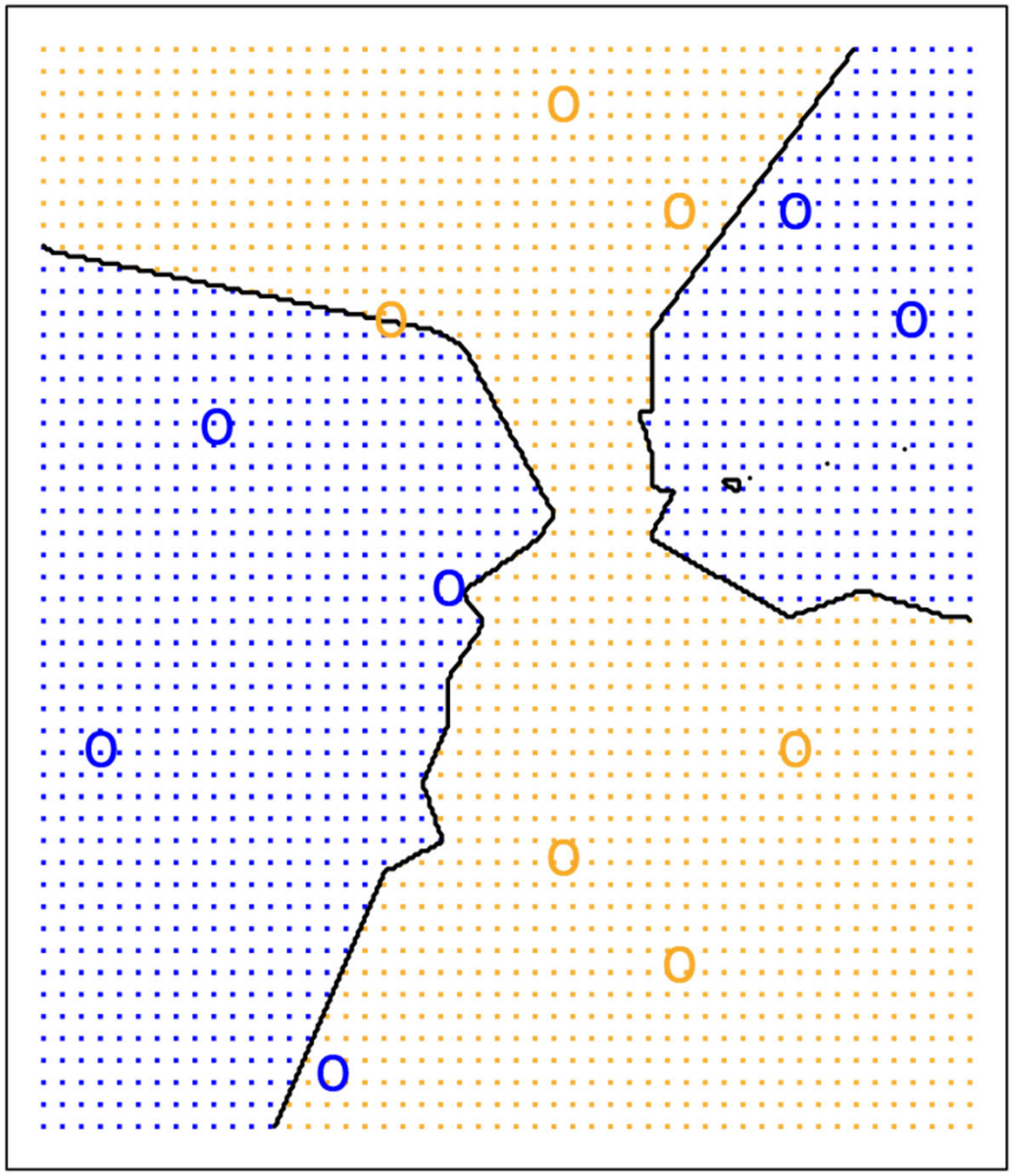
- 왜 필요한가? 베이즈 분류기에서 조건부 확률 $P(Y|X)$을 계산할 수 없을 때
- $P(Y|X)$를 추정하기 위한 한 방법으로 KNN!

$$P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j)$$

  - K: 주어진 양의 상수
  - $\mathcal{N}_0$: $X = x_0$에 근접해 있는 K개 훈련세트 개체들의 집합
- 예: 가상 자료(그림2.14)
  - $K = 3$
  - K-최근접 이웃 결정 경계(KNN decision boundary): 검정색 실선

그림2-14 가상 자료.
K=3. 왼쪽(x: 테스트 개체, 파란색 범주로 분류), 오른쪽(검은색: KNN 결정 경계)





K-최근접 이웃 분류기

- 예: 가상 자료(그림2.15, 2.16)
  - $K = 10$
  - 테스트 오류율: 0.1363(KNN)
  - $K = 1$ vs. 100 일 때: (과도하게 유연한(overly flexible) ⇒ 낮은 편의와 높은 분산) vs. (결정 경계가 거의 선형적 ⇒ 높은 편이와 낮은 분산)
  - 테스트 오류율은 각각 0.1695, 0.1925

K-최근접 이웃 분류기

- 회귀 문제처럼 유연성이 커지면 훈련 오류율은 작아지지만 테스트 오류율은 “U”자 형태를 보임
- 5장에서 테스트 오류율을 추정하는 여러 가지 방법을 다룸!

그림2-15 보라색 베이즈 결정 경계, 검은색: KNN(K=10) 결정 경계

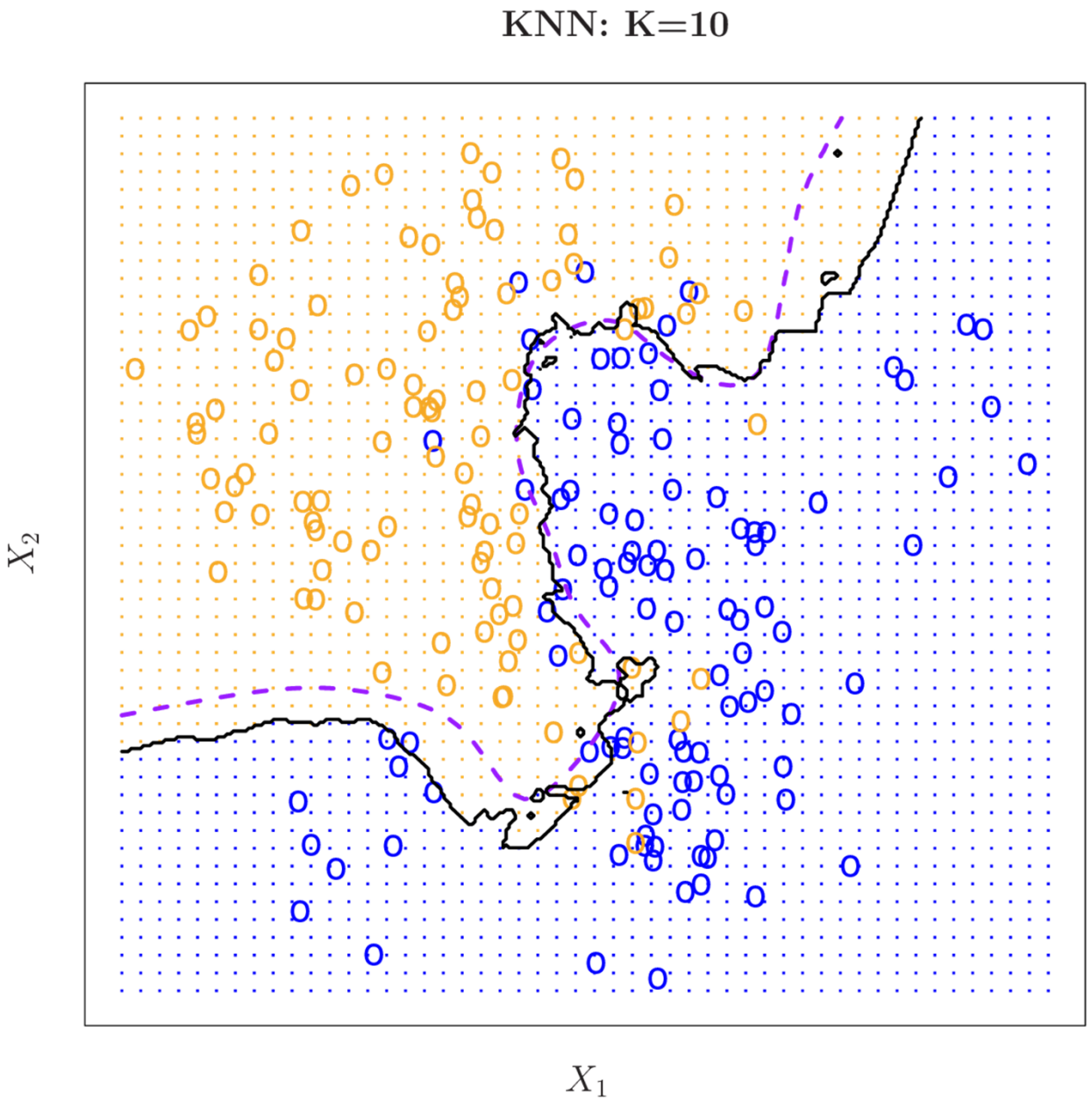


그림2-16 보라색: 베이즈 결정 경계, 검은색: KNN 결정 경계

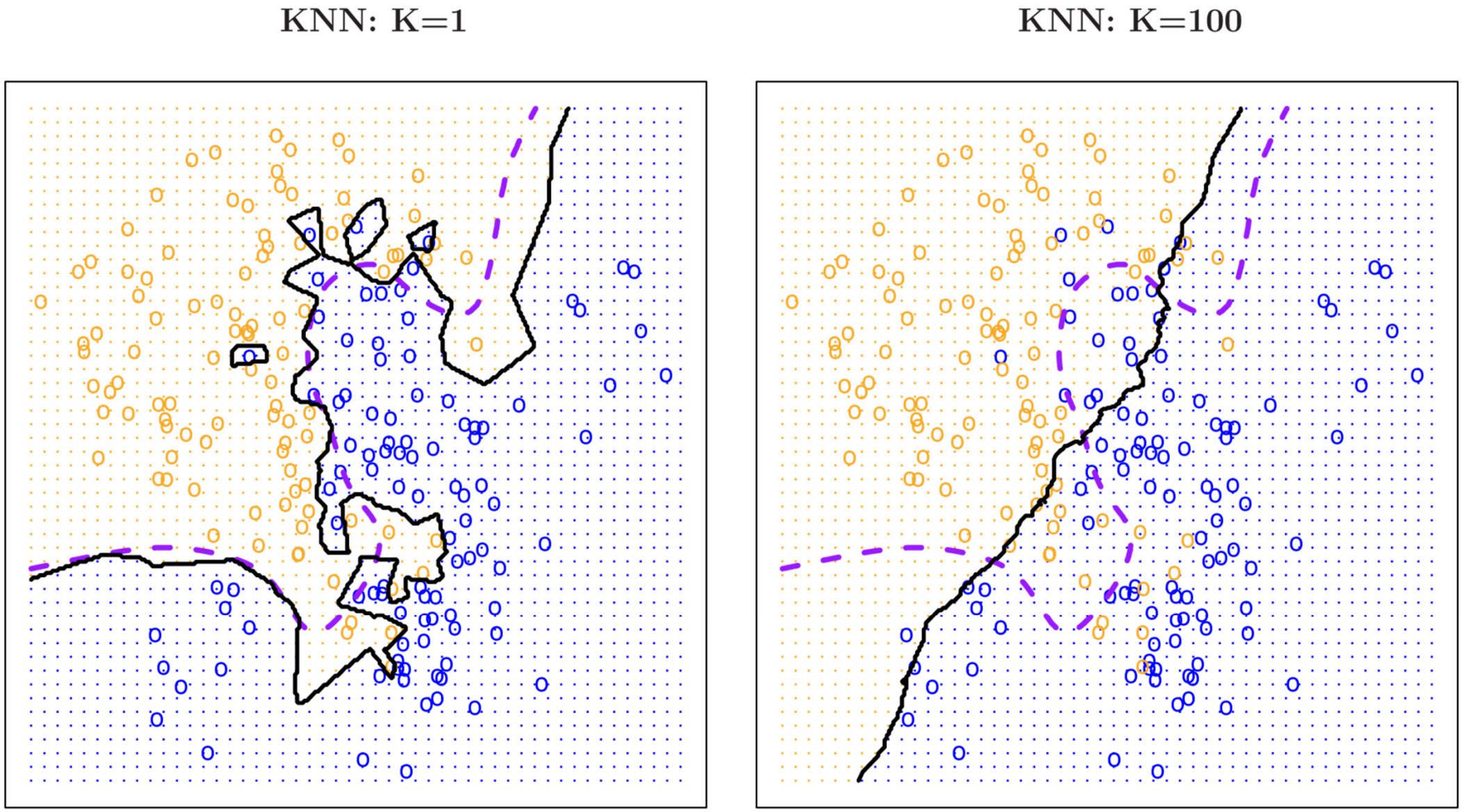
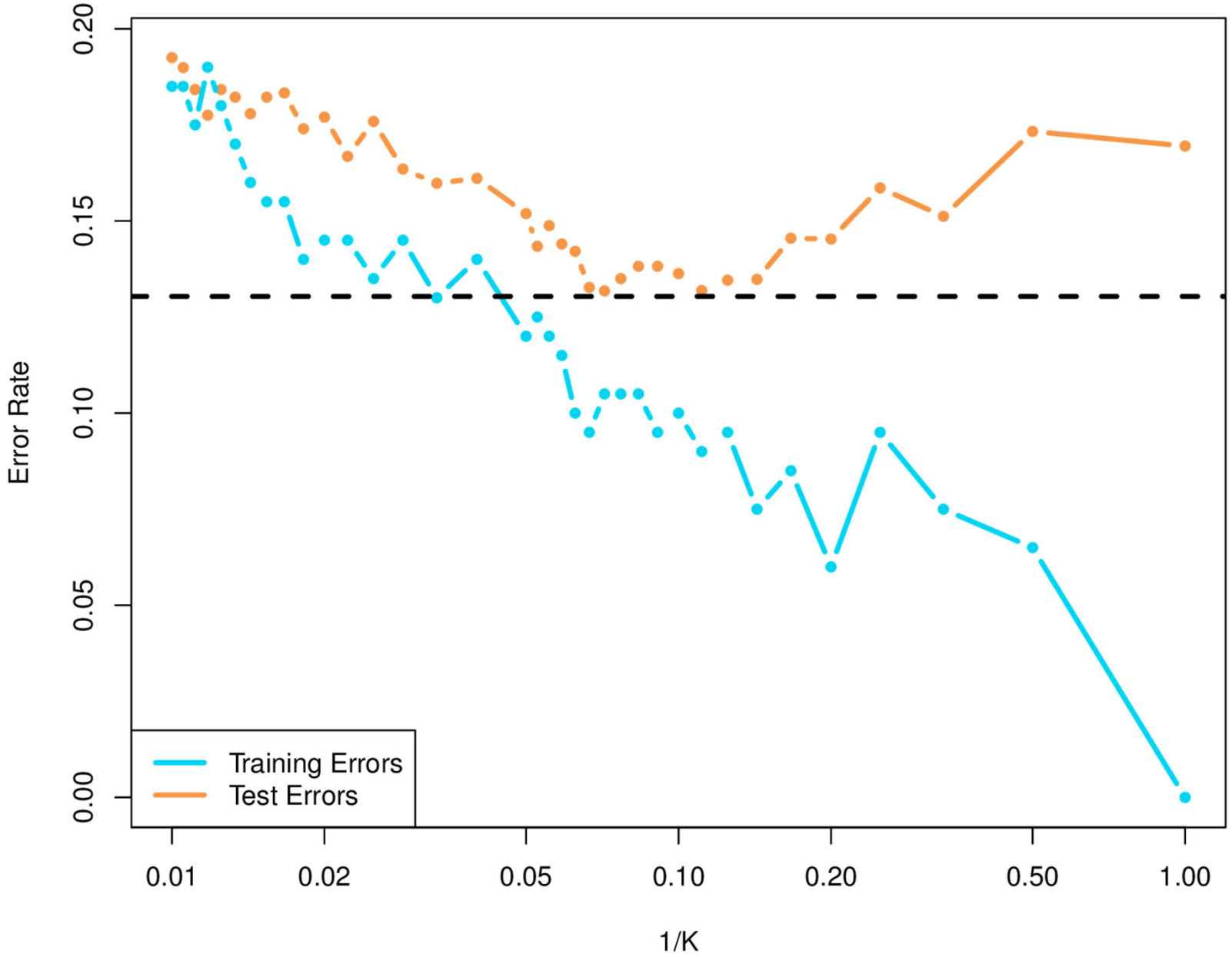


그림2-17
검은색:
베이즈
오류율,
파란색:
훈련(n=20)
오류율,
오렌지색:
테스트(n=
5000)
오류율



과제(4월7일마감)

- 연습문제: 1, 2, 6, 10

Thank you!

Move on to 03 Linear regression

03 Linear regression

J Kim

2021.3

Outline

단순 선형회귀모형
다중 선형회귀모형
그 외 고려할 점들
KNN과 비교

단순 선형회귀모형

- 지도학습 방법 중 간단. 그러나 기본!
- 목표: 양적 반응변수를 예측하는 것
- 가정: $Y = \beta_0 + \beta_1 X + \epsilon$
  - β_0: 절편(intercept)
  - β_1: 기울기(slope)
  - β_0, β_1: 미지의(unknown) 회귀계수(regression coefficients) 혹은 회귀모수(regression parameters)라고 함
  - ϵ: 랜덤 오차, $E(\epsilon) = 0$

회귀계수 추정

- (훈련)데이터 세트: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- 목표: 데이터 세트에 가장 가까운 직선을 얻는 것. 즉 $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$
- 해법: 최소제곱(least squares) 원리! 즉, 잔차제곱합(residual sum of squares; RSS)을 가장 작게 하는 값을 $\hat{\beta}_0, \hat{\beta}_1$으로!
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: 예측값(predicted value)
  - $e_i = y_i - \hat{y}_i$: 잔차(residual)
  - RSS= $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$
  - LSE(least squares estimates):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Remarks:

- $f(X) = \beta_0 + \beta_1 X$: True(모)회귀직선(population regression line)
- $\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$: 최소제곱 회귀직선(least squares regression line)
- LS 직선들의 평균 $\approx$ 모회귀직선

그 림 3-7:
광고 자료

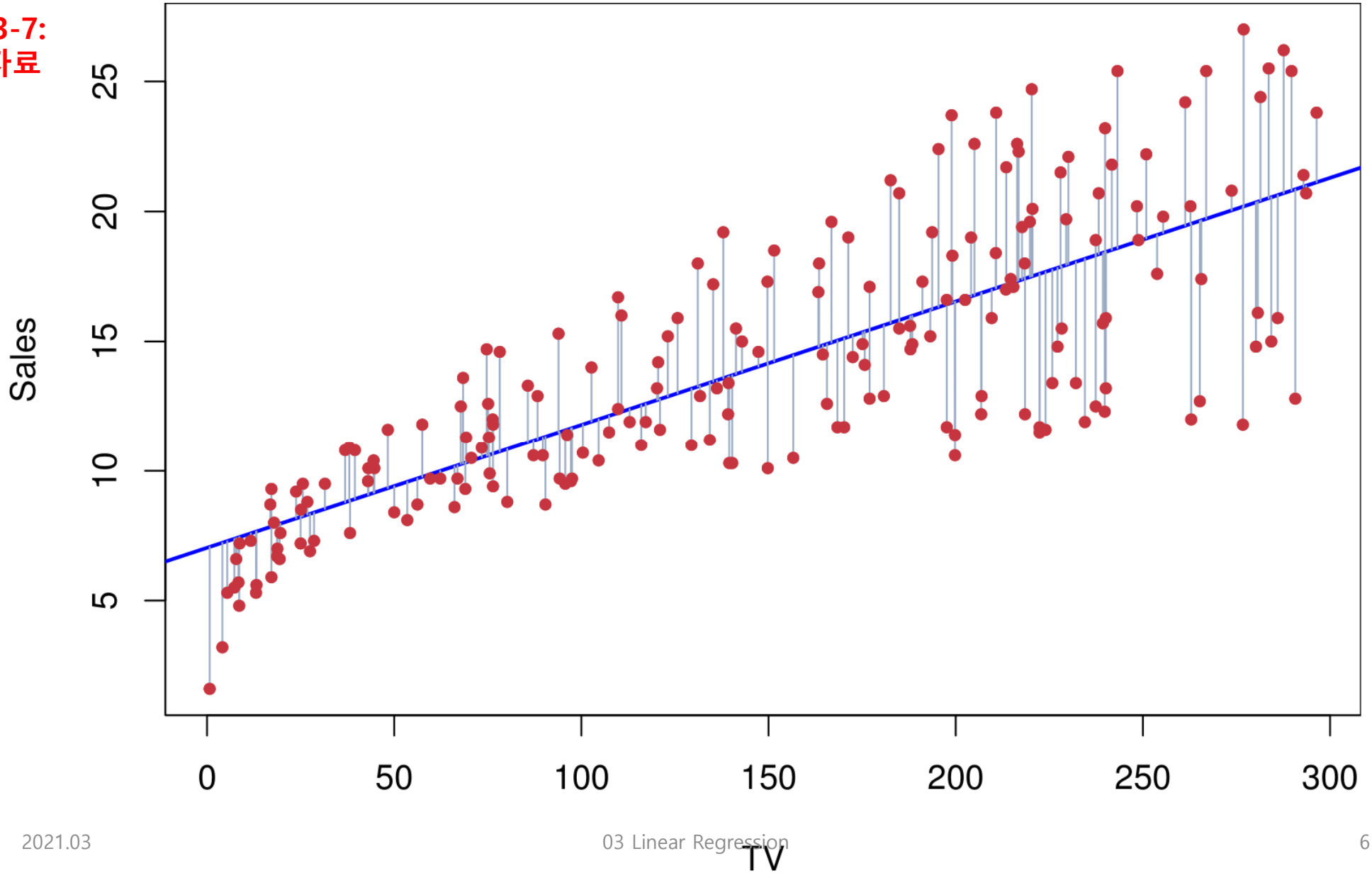
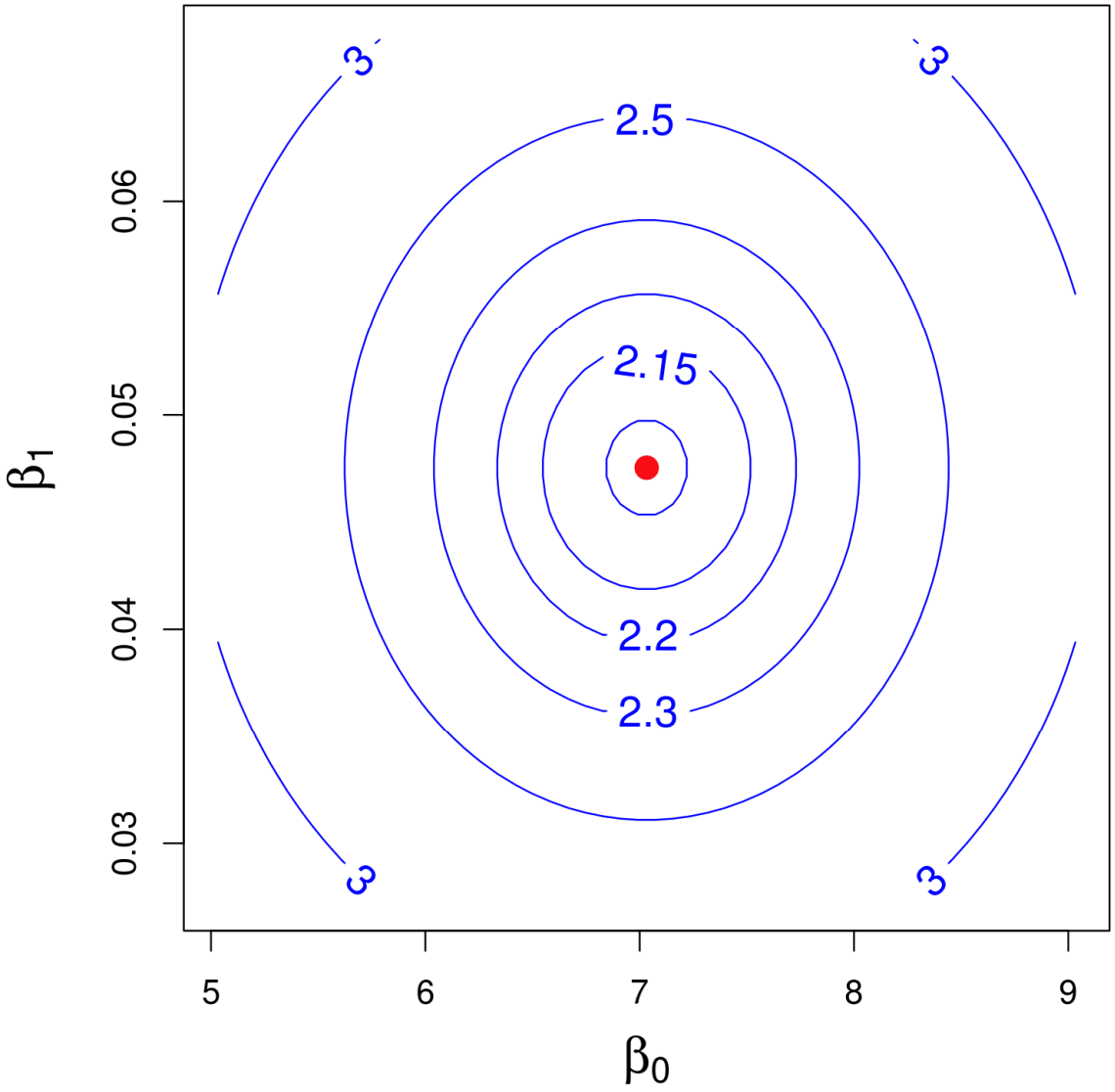


그림3-2: 광고 자료. 왼쪽(RSS 등고선), 오른쪽(RSS)



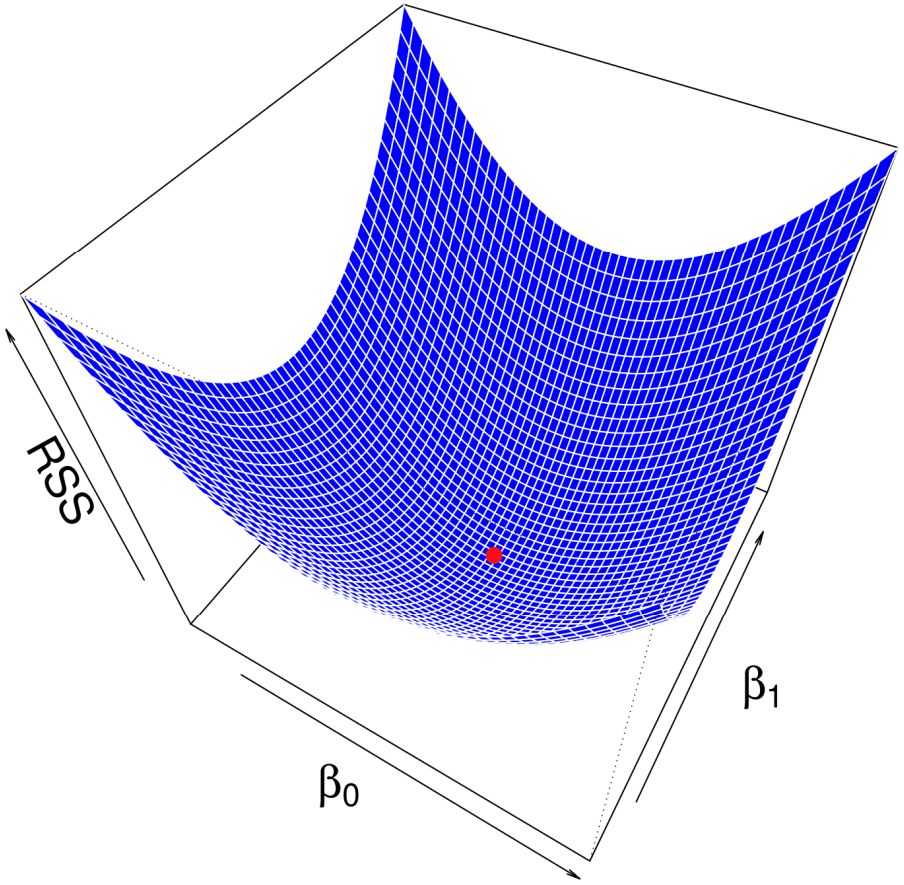
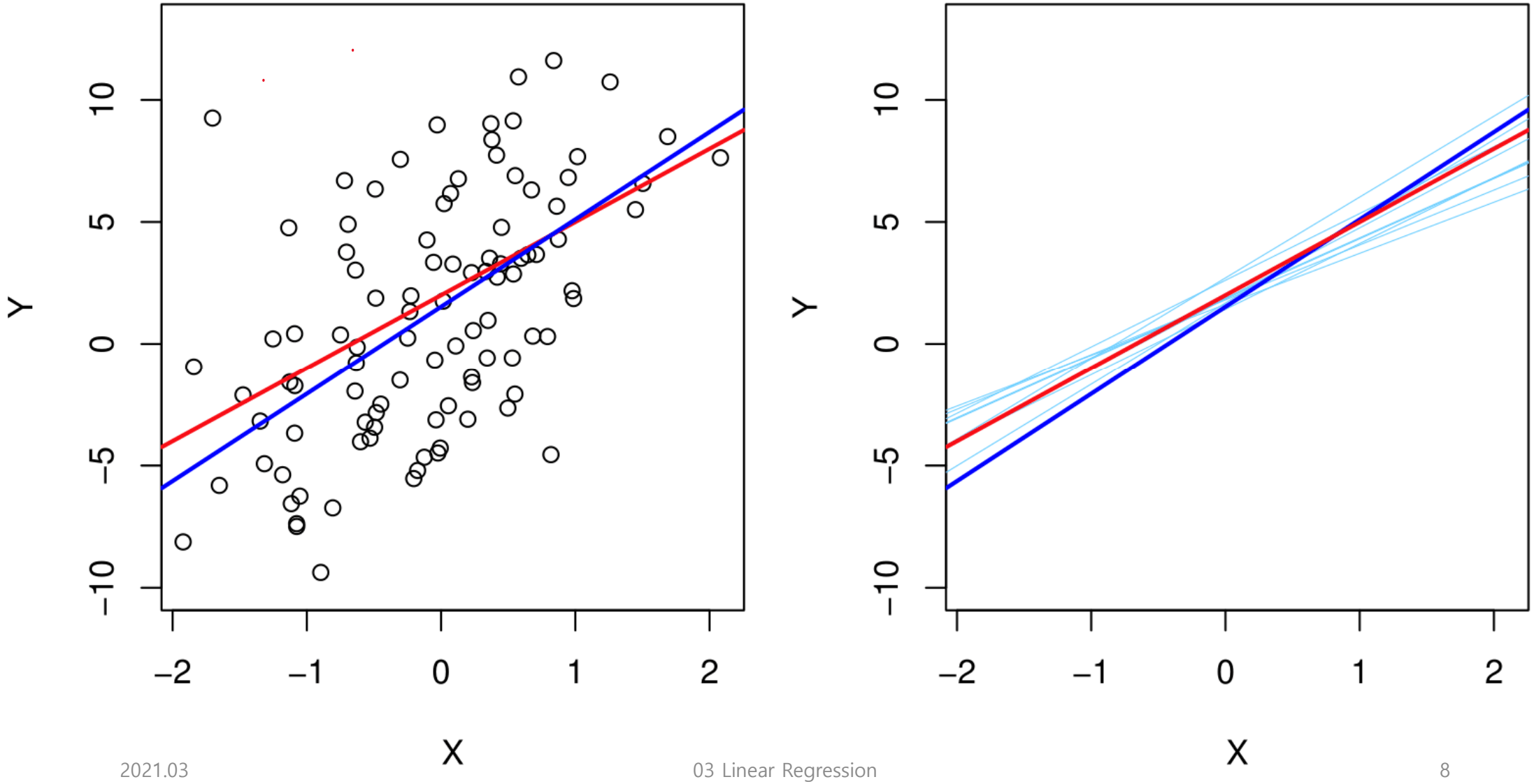


그림3-3: 가상 자료. f(X)=2+3X. 빨간색(True 회귀직선), 파란색(LS 회귀직선)



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars).*

회귀추정량은 얼마나 정확한가?

- 불편(unbiased)추정량: $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$
- 표준오차(standard error):

$$\mathrm{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad \mathrm{SE}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- 추가 가정: $Var(\epsilon) = \sigma^2$, $Cov(\epsilon_i, \epsilon_j) = 0 (i \neq j)$

Remarks:

- $x_i (i = 1,2, \dots, n)$가 넓게 퍼져 있을수록 표준오차가 줄어듦
- σ^2의 추정: $\hat{\sigma}$ =RSE=$\sqrt{\frac{\mathrm{RSS}}{n-2}}$: 잔차표준오차(residual standard error)
- $\hat{\sigma}$를 대입하면 $\widehat{\mathrm{SE}}(\hat{\beta}_0), \widehat{\mathrm{SE}}(\hat{\beta}_1)$로 표현해야 하지만 표현을 간단히!

95% 신뢰구간추정

- $\beta_0 \in \left[\hat{\beta}_0 - t_{0.025}(n-2)\mathrm{SE}\left(\hat{\beta}_0\right), \hat{\beta}_0 + t_{0.025}(n-2)\mathrm{SE}\left(\hat{\beta}_0\right)\right]$
- $\beta_1 \in \left[\hat{\beta}_1 - t_{0.025}(n-2)\mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + t_{0.025}(n-2)\mathrm{SE}\left(\hat{\beta}_1\right)\right]$
- 추가 가정: $\epsilon_i (i = 1,2,\ldots,n)$: 정규분포를 따름

가설검정

- $H_0: \beta_1 = 0$ 대 $H_a: \beta_1 \neq 0$
- "X와 Y는 관계가 없는지"를 검정
- t-통계량: $t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
  - "$\hat{\beta}_1$이 0에서 표준오차의 몇 배 떨어져 있는지에 대한 값"을 의미
- p-값 $= P(|t| \geq |t^*|), t \sim t(n-2)$
  - t^*: t-통계량의 관측값
- p-값이 0.05보다 작으면 H_0를 기각 (X와 Y는 관계 있음!)

모형은 얼마나 정확한가?
: RSE

- 모형에 랜덤 오차(ϵ)가 있기 때문에 True 회귀직선을 안다고 해도 X를 써서 Y를 완전하게(perfectly) 예측할 수 없음!
- ϵ의 표준편차의 추정량
- 대략 말해 "평균적으로 Y가 True 회귀직선에서 벗어난 양"으로 해석
- RSE가 작을수록 모형이 데이터에 잘 적합됨!

R^2

- 정의: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$
- TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$: 총제곱합(total sum of squares) (회귀모형과 무관!)
- TSS – RSS: 모형에 의해 설명되는 Y 의 변동
- "X를 써서 설명되는 Y의 변동의 비율"을 의미!
- $R^2 \in (0,1) \Rightarrow$ "$R^2 \approx 1$"이면 모형이 데이터에 잘 적합됨!
- Remarks: $R^2 = r^2$ WHY?
  - $r = Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$: 표본상관계수

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

다중 선형회귀모형:
왜 다중회귀모형인가?

- 2개 이상 예측변수가 주어졌을 때 단순회귀모형으로는 어떻게 예측해야 할지 명확하지 않음
- 한 예측변수로 된 단순회귀모형으로 예측하면 나머지 예측변수는 무시하는 꼴!
- 예측변수들이 상관되어(correlated) 있으면 한 예측변수로 된 단순회귀모형은 잘못된 추정이 될 수도!

왜 다중회귀모형인가?

- 예: 상어 공격과 아이스크림 판매량의 관계 ⇒ 양의 관계!
  그러나 기온을 포함하면 아무 관계 없음.
- 왜? 기온 상승 ⇒ 사람들이 비치로 운집 ⇒ 아이스크림 판매 ↑ &
  상어 공격 ↑

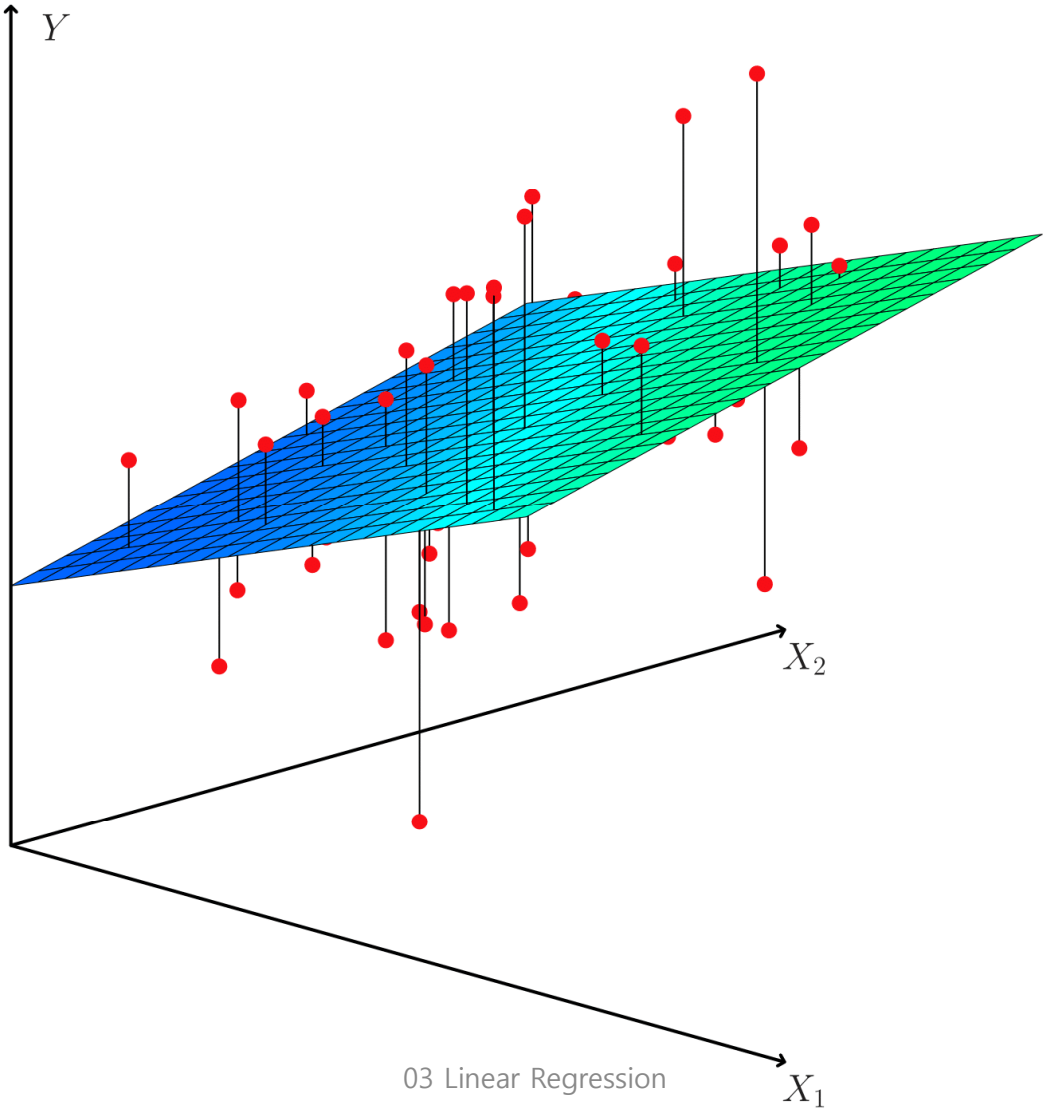
회귀계수 추정

- 모형: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$
  - $\beta_0, \beta_1, \ldots, \beta_p$: 미지의 회귀계수
  - "$\beta_j (j = 1,2, \ldots, p)$ " 의 의미: X_j를 제외한 나머지 예측변수는 고정시킨 후 X_j의 한 단위(unit) 변화가 Y에 미치는 평균적인 효과의 크기
- (훈련)데이터 세트:
  - $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$
- LSE: "RSS = $\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2$"을 가장 작게 하는 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$으로!
- 예측값: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, i = 1,2, \ldots, n$

Remarks:

- $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$: True(모)회귀평면(population regression plane)
- $\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \cdots + \hat{\beta}_p X_p$: 최소제곱 회귀평면(LS regression plane)

그림3-4: 가상 자료. 파란색 평면(LS 회귀평면)



Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

TABLE 3.3. *More simple linear regression models for the* `Advertising` *data. Coefficients of the simple linear regression model for number of units sold on* Top: *radio advertising budget and* Bottom: *newspaper advertising budget. A* \$1,000 *increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units (Note that the* `sales` *variable is in thousands of units, and the* `radio` *and* `newspaper` *variables are in thousands of dollars).*

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

광고자료: 신문 광고와 매출은 관계가 없는가?

- 맞음!
- 단순 선형회귀모형에서는 신문 광고가 매출을 증가시킨다고 했는데 …
- 라디오 광고와 신문광고의 상관계수가 0.3541임. 즉 라디오 광고를 많이 하면 신문 광도도 많이 하는 경향이 있음
- 라디오 광고가 많을수록 매출은 증가하고 라디오 광고가 많을수록 신문광고도 증가하므로 단순 선형회귀모형과 같은 결과가 나옴
- 신문 광고에 의한 매출은 라디오 광고에 대한 대리(surrogate)효과라고 할 수 있음

예측변수와 반응변수는 관계가 있는가? 가설검정

- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 대 H_a: $\beta_j (j = 1,2, \ldots, p)$ 중 적어도 한 개는 0이 아니다
- F-통계량: $F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$
- p-값 $= P(F \geq F^*), F \sim F(p, n-p-1)$
  - F^*: F-통계량의 관측값

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

TABLE 3.6. *More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the* `Advertising` *data. Other information about this model was displayed in Table 3.4.*

부분(partial) F-검정

- 예측변수 $q(\leq p)$개로 된 부분집합에 대한 가설검정
  - $H_0: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ 대 H_a: $\beta_j (j = p-q+1, p-q+2, \ldots, p)$ 중 적어도 한 개는 0이 아니다
  - F-통계량: $F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$
  - RSS_0: 부분집합에 포함되지 않은 예측변수로 된 모형의 RSS
  - P-값 $= P(F \geq F^*), F \sim F(q, n-p-1)$
- Remarks: 만약 $q = 1$이면 개별 예측변수에 대한 검정과 동일! 사실 $F = t^2$!

변수선택

- 모든 가능한 모형은 2^p개!
- 측도: Mallow의 C_p, AIC(Akaike information criterion), BIC(Bayesian information criterion), 수정된(adjusted) R^2 등
- p가 크면 가능한 모형이 매우 많음

전진선택(forward selection)

- 절편항 모형(null model)에서 시작
- null 모형에 예측변수 한 개를 추가한 모형을 p개 적합한 후 RSS가 가장 작은 모형을 선택
- 정지 규칙이 만족될 때까지 진행
- 탐욕적인(greedy) 방법

후진선택(backward selection)

- 모든 예측변수를 포함한 모형(full model)에서 시작
- 가장 p-값이 큰 예측변수를 제거한 후 나머지 $(p-1)$개 예측변수만으로 모형을 다시 적합
- 정지 규칙이 만족될 때까지 진행
- "$p > n$"인 자료에는 적용할 수 없음!
- Remarks: 혼합선택(mixed selection)
  - 전진선택 방법과 동일하게 진행하되 p-값이 임계값(threshold)보다 큰 예측변수는 모형에서 제거

모형 적합(model fit)

- 예측변수가 반응변수와 관계가 약할지라도 모형에 포함시키면 R^2는 증가함! 증가량이 둔화될 때
  - 광고 자료: TV만 (0.612) → TV + radio (0.89719) → TV + radio + newspaper (0.8972)
- Remarks: $R^2 = Cor(Y, \hat{Y})^2$ WHY?
- RSE: RSE = $\sqrt{\frac{\text{RSS}}{n-p-1}}$가 작을수록
  - 광고 자료: TV만 (3.26) → TV + radio (1.681) → TV + radio + newspaper (1.686)

신뢰구간(confidence interval)과 예측구간(prediction interval_

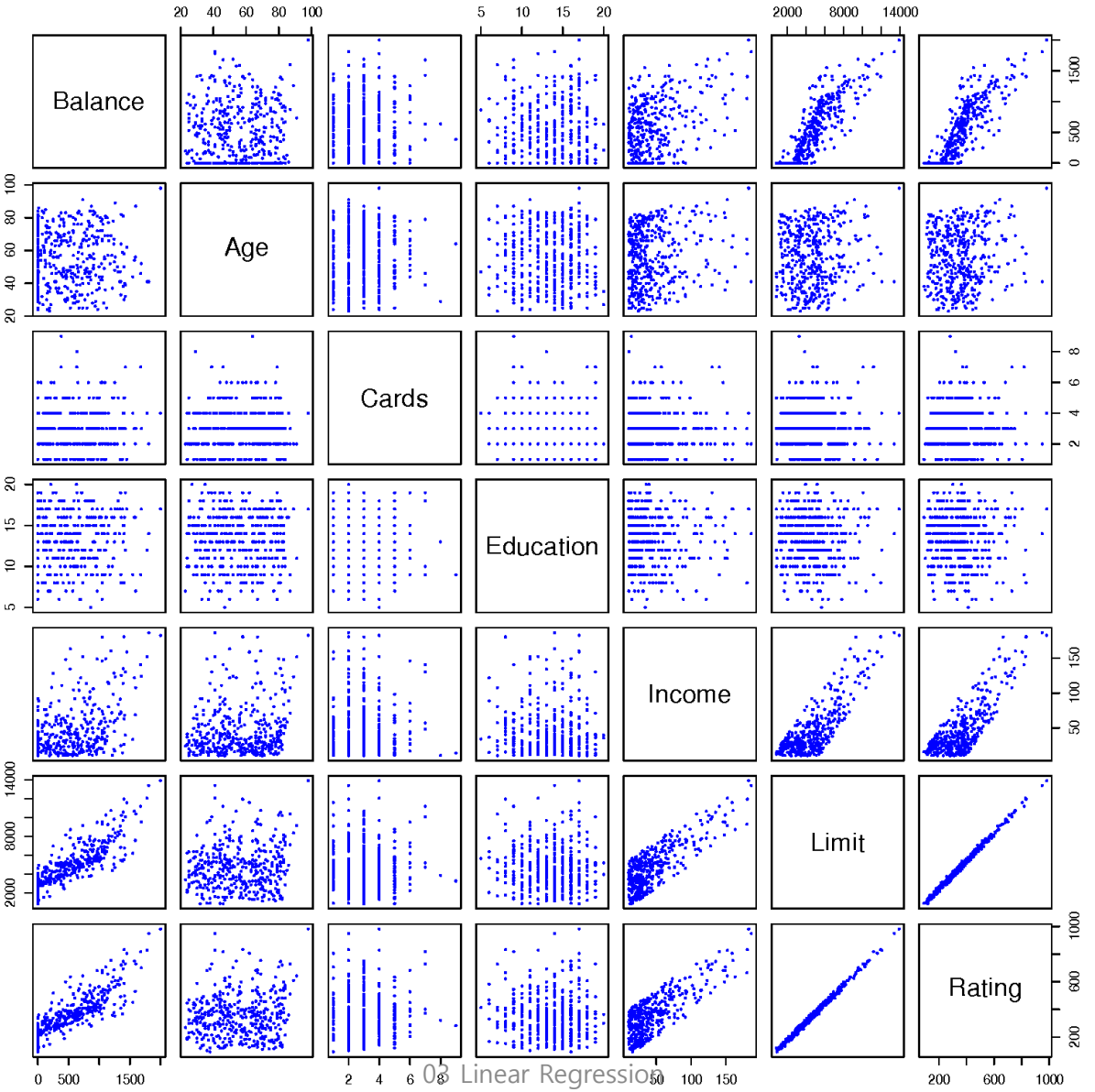
- 반응변수의 평균적인 값에 대한 신뢰구간 계산!
  - 광고 자료: TV = 10만불, radio = 2만불일 때 평균 매출에 대한 95% 신뢰구간 = $(10985, 11528)$
  - "평균 매출에 대한 신뢰구간을 의미"
- 특정 개체의 반응변수의 값에 대한 예측구간 계산
  - 신뢰구간보다 폭이 넓어짐 WHY? ϵ
  - 광고 자료 : TV = 10만불, radio = 2만불일 때 매출에 대한 95% 예측구간 = $(7930, 14580)$
  - "특정 도시의 매출에 대한 예측구간을 의미"

질적 예측변수

- 예: 신용(credit) 자료-고객 400명에 대한 가상 자료
  - 어떤 고객이 신용카드 채무 불이행을 범할까?
  - 반응변수: 신용카드 빚
  - 예측변수: 나이, 신용카드수, 교육연수, 수입, 한도, 신용등급 (이상 양적)
    성별, 학생 여부, 혼인 여부, 민족 (이상 질적)

ID Identification
Income Income in $10,000's
Limit Credit limit
Rating Credit rating
Cards Number of credit cards
Age Age in years
Education Number of years of education
Gender A factor with levels Male and Female
Student A factor with levels No and Yes
Married A factor with levels No and Yes
Ethnicity A factor with levels African American, Asian, and Caucasian
Balance Average credit card balance in $.

그림3-7: 신용 자료



질적 예측변수:
범주가 2개일 때

- 예측변수: 성별
- 지시(indicator) 혹은 더미(dummy) 변수를 1개 생성
- $x_i = \begin{cases} 1, \text{여자} \\ 0, \text{남자} \end{cases}$ 혹은 $x_i = \begin{cases} 0, \text{여자} \\ 1, \text{남자} \end{cases}$ 혹은 $x_i = \begin{cases} 1, \text{여자} \\ -1, \text{남자} \end{cases}$
- 회귀모형: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i =$
  - $\begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{여자} \\ \beta_0 + \epsilon_i, & \text{남자} \end{cases}$ 혹은 $\begin{cases} \beta_0 + \epsilon_i, & \text{여자} \\ \beta_0 + \beta_1 + \epsilon_i, & \text{남자} \end{cases}$ 혹은 $\begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{여자} \\ \beta_0 - \beta_1 + \epsilon_i, & \text{남자} \end{cases}$
- Remarks: 예측값은 세 가지 코드 방법 모두 동일!

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* balance *onto* gender *in the* Credit *data set. The linear model is given in (3.27). That is, gender is encoded as a dummy variable, as in (3.26).*

범주가 3개일 때

- 예측변수: 민족
- 더미 변수를 2개 생성
- $x_{i1} = \begin{cases} 1, & \text{아시안} \\ 0. & \text{아시아안 아님}' \end{cases}$ $x_{i2} = \begin{cases} 1, & \text{코카서스인} \\ 0, & \text{코카서스인 아님} \end{cases}$
- 회귀모형
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i, & \text{아프리카계 미국인} \\ \beta_0 + \beta_1 + \epsilon_i, & \text{아시안} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{코카서스인} \end{cases}$

Remarks:

- 개별 가설검정($H_0: \beta_1 = 0; H_0: \beta_2 = 0$)은 무의미하고 가설 $H_0: \beta_1 = \beta_2 = 0$에 대한 검정이 필요! 검정결과는 코드 방법에 무관하게 동일!

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	−18.69	65.02	−0.287	0.7740
ethnicity[Caucasian]	−12.50	56.68	−0.221	0.8260

TABLE 3.8. *Least squares coefficient estimates associated with the regression of* balance *onto* ethnicity *in the* Credit *data set. The linear model is given in (3.30). That is, ethnicity is encoded via two dummy variables (3.28) and (3.29).*

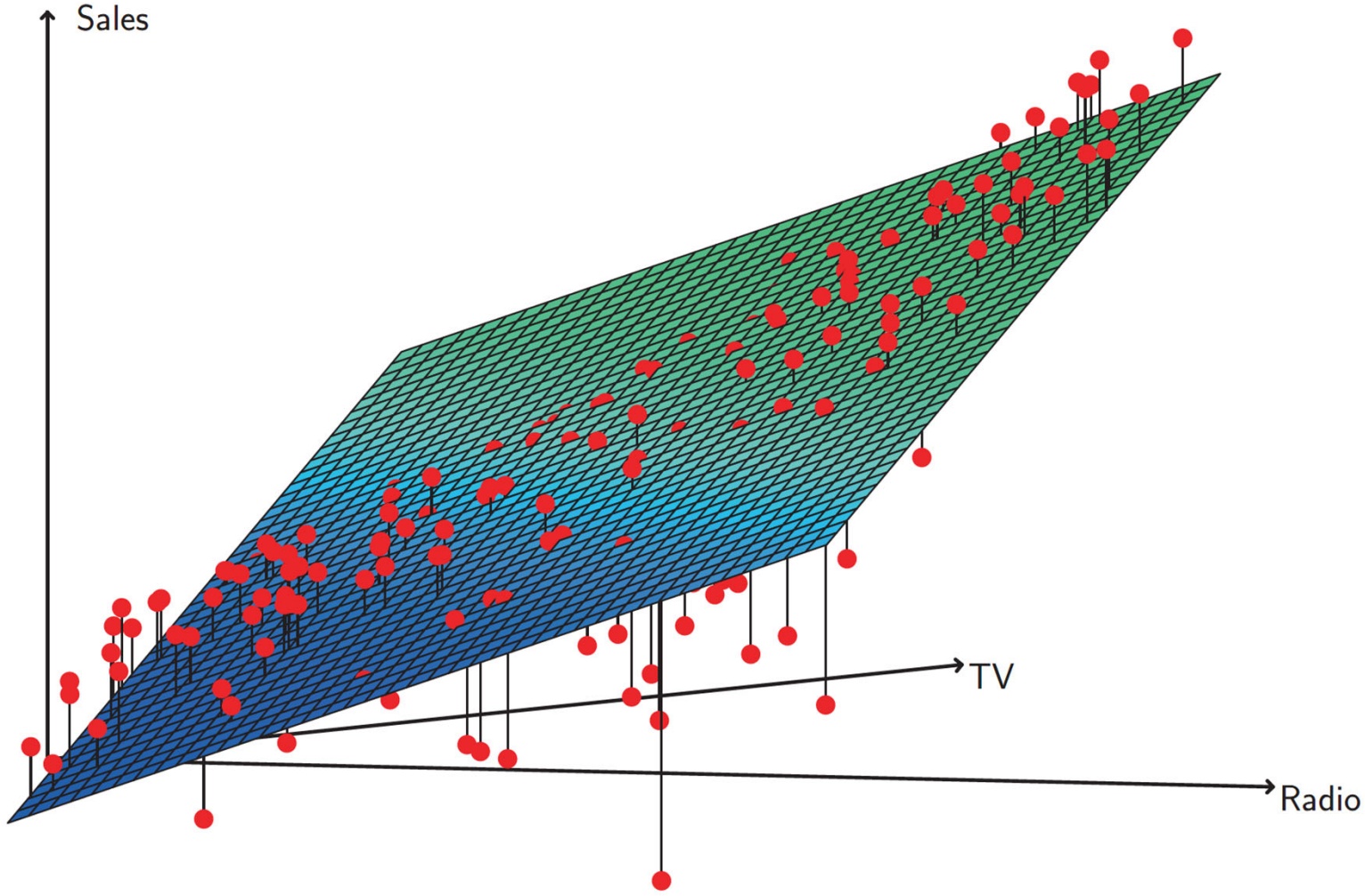
선형모형 확장

- 선형적(linear)이고 가법적(additive)인 모형
- 가법적이란? 한 예측변수의 변화가 반응변수에 미치는 효과 다른 예측변수와 무관하다는 뜻
- 선형적이란? 한 예측변수의 변화가 반응변수에 미치는 효과가 그 예측변수의 값에 무관하게 일정하다는 뜻

가법성 가정 제거: 광고 자료

- TV 혹은 라디오 중 한 매체에 집중하면 과대추정(overestimate)하고, 두 매체로 분할하면 과소추정(underestimate)!
- 시너지 효과 혹은 상호작용 효과 존재 시사!

그림3-5: 광고 자료. 양의 잔차(과소 추정)는 광고가 골고루, 음의 잔차(과대 추정)는 광고가 한 매체로 치우침



상호작용효과 모형

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

  $= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$

  $= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon, \tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

  혹은

  $Y = \beta_0 + \beta_1 X_1 + \tilde{\beta}_2 X_2 + \epsilon, \tilde{\beta}_2 = \beta_2 + \beta_3 X_1$

- $\tilde{\beta}_1$가 X_2에 따라 변하기 때문에 X_1이 Y에 미치는 영향이 더 이상 X_2의 변화와 무관하지 않음!

상호작용효과 모형

- β_3: X_2(혹은 X_1)를 한 단위 증가 시킴으로 인해 X_1 (혹은 X_2)이 Y에 미치는 효과의 양으로 해석
- $E(Y|X_1 = a, X_2 = b+1) = \beta_0 + \beta_1 a + \beta_2(b+1) + \beta_3 a(b+1)$

  $= (\beta_0+\beta_2) + \beta_1 a + \beta_2 b + \beta_3 ab + \beta_3 a$

  $E(Y|X_1 = a, X_2 = b) = \beta_0 + \beta_1 a + \beta_2 b + \beta_3 ab$

  $\therefore E(Y|X_1 = a, X_2 = b+1) - E(Y|X_1 = a, X_2 = b) = \beta_2 + \beta_3 a$
- 즉 β_3에도 의존함. $\beta_3 > 0$이면 X_1이 시너지 효과를 내며 그 크기는 X_1의 값에 따라 변함!

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

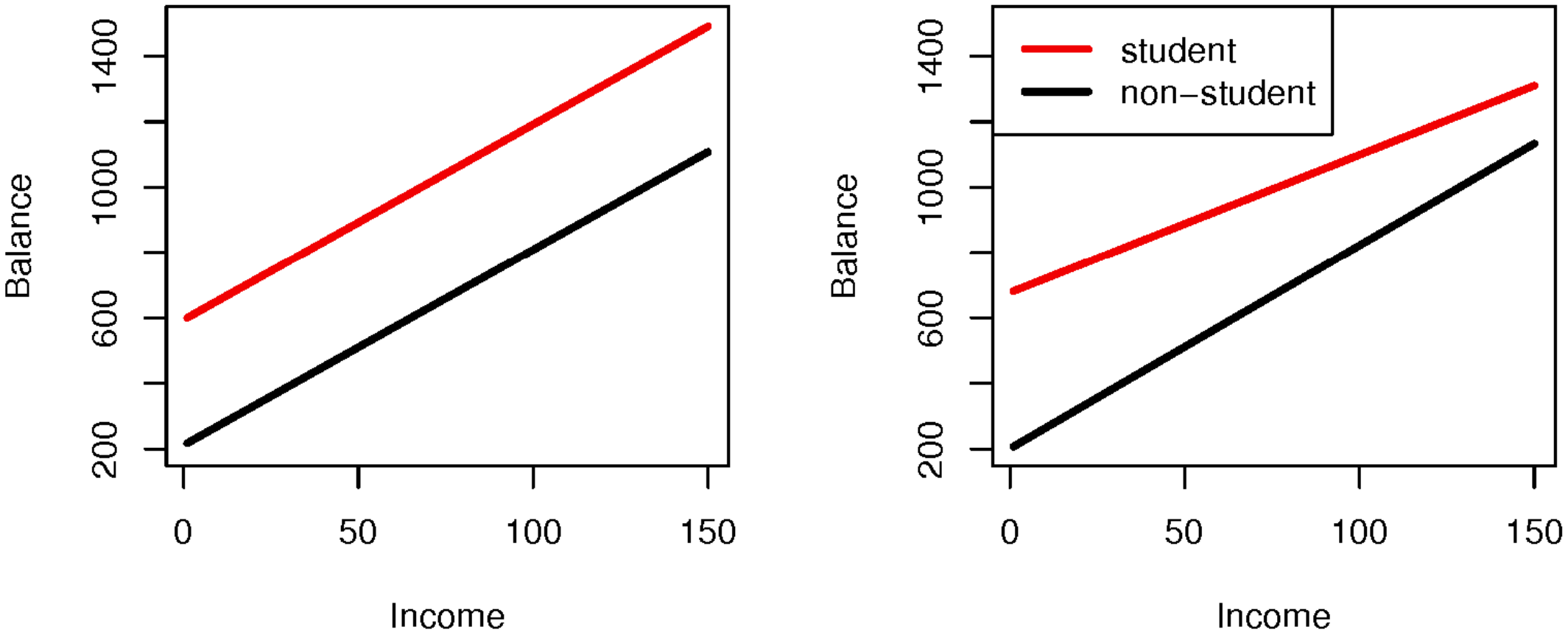
가법성 검정

- $H_0: \beta_3 = 0$ 대 $H_a: \beta_3 \neq 0$
- Remarks: 계보적(hierarchical) 모형에서는 상호작용이 모형에 포함되면 주효과(main effect)가 통계적으로 유의하지 않더라도 모형에 포함되어야 한다는 원리

Remarks:

- 질적변수 간의 상호작용, 질적변수와 양적변수 간의 상호작용도 동일하게 다룸!
- 예: 신용 자료 - 수입과 학생 여부 간의 상호작용 효과 포함
  - 상호작용 효과 포함하지 않으면: 기울기는 동일하나 절편항은 다름
  - 상호작용 효과 포함하면: 기울기와 절편항 모두 다름

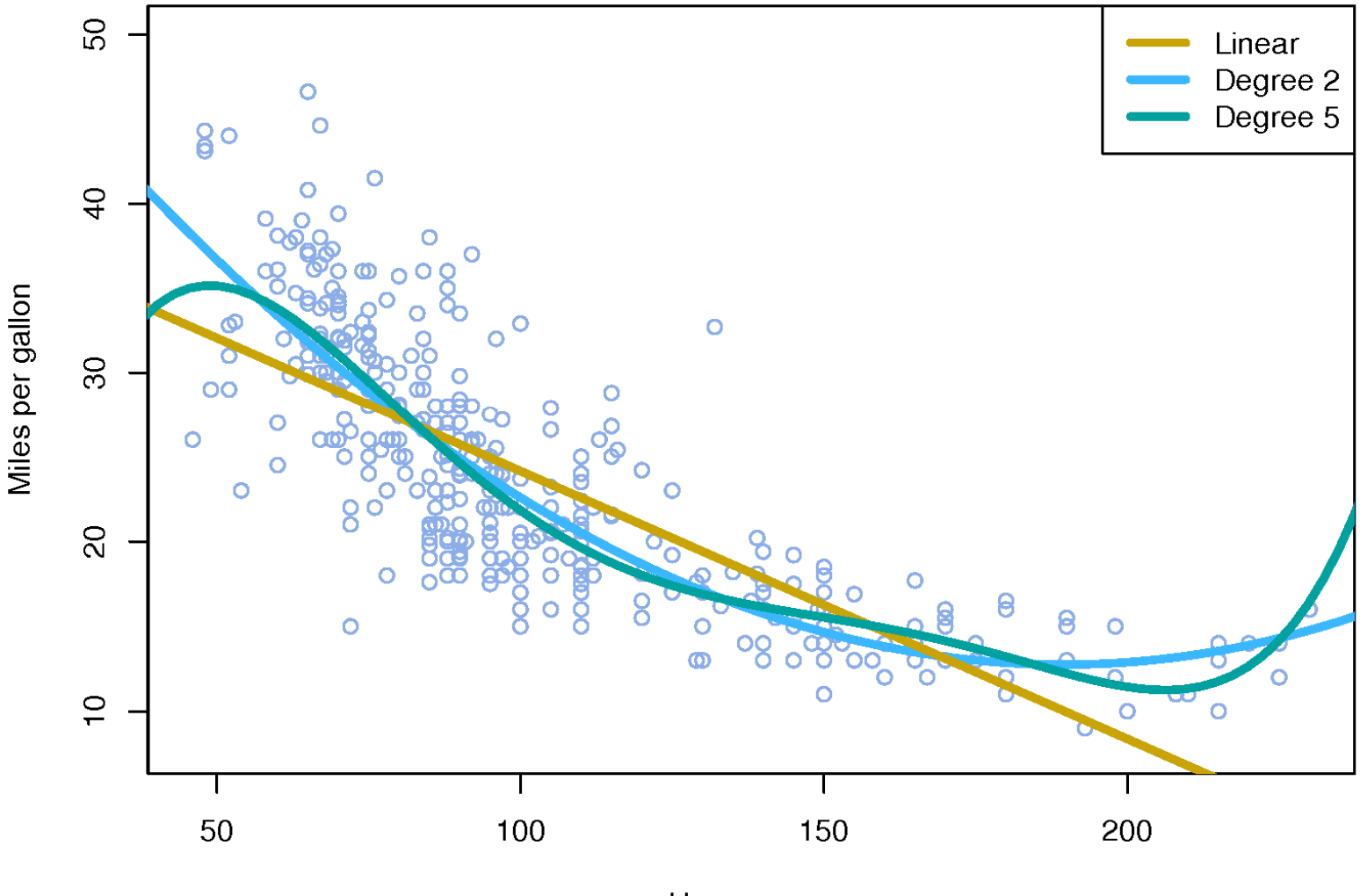
그림3-7: 신용 자료. 왼쪽(수입과 학생 유무 간 상호작용효과 없음), 오른쪽(상호작용효과 있음)



비선형성

- 간단한 해법: 다항회귀모형 - 예측변수의 거듭제곱을 모형에 포함
- 그러나 여전히 선형모형!
- 예: 자동차 자료
  - 반응변수: 연비
  - 예측변수: 마력
  - 이차 다항회귀모형이 가장 우수!

그림3-8: 자동차 자료


Linear
Degree 2
Degree 5
Miles per gallon
10
20
30
40
50
50
100
150
200
Horsepower

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

TABLE 3.10. *For the* Auto *data set, least squares coefficient estimates associated with the regression of* mpg *onto* horsepower *and* horsepower^2.

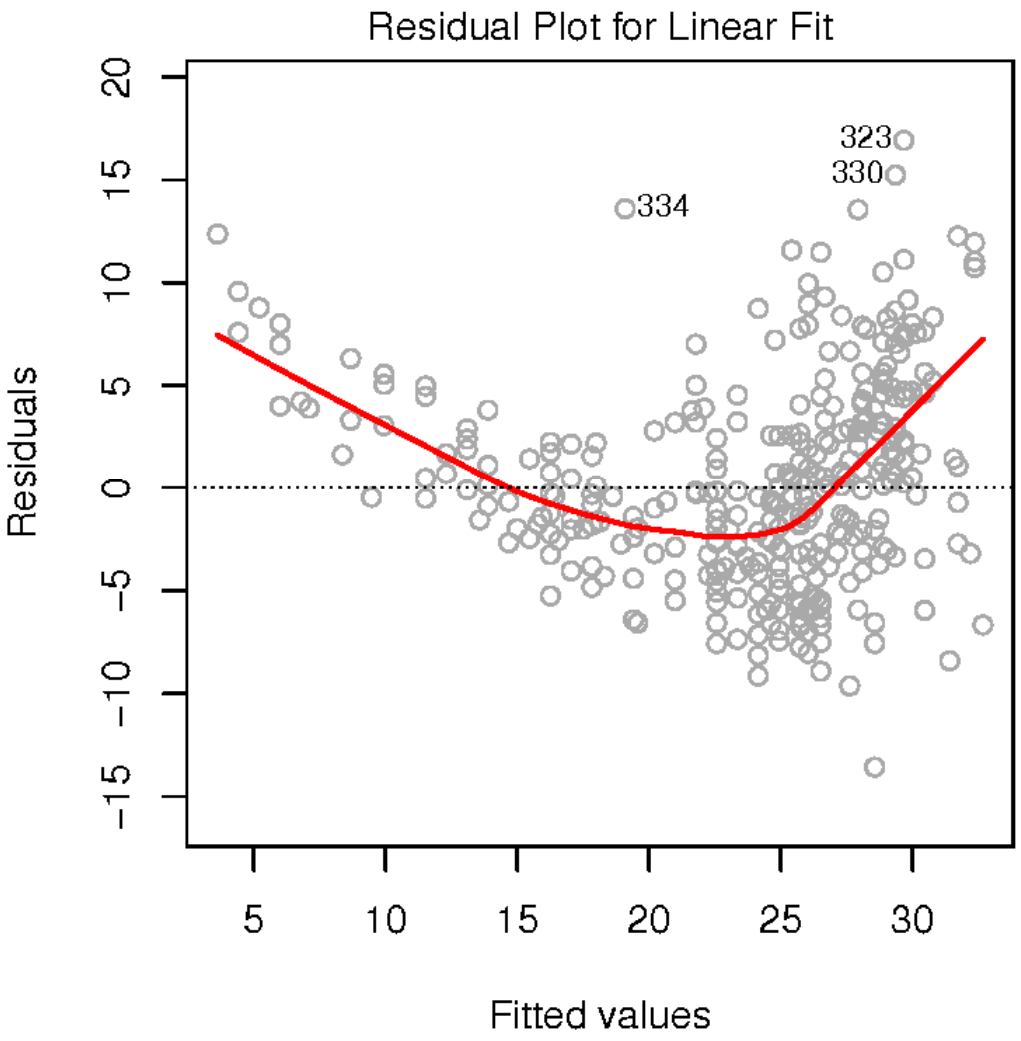
그 외 가능한 문제들

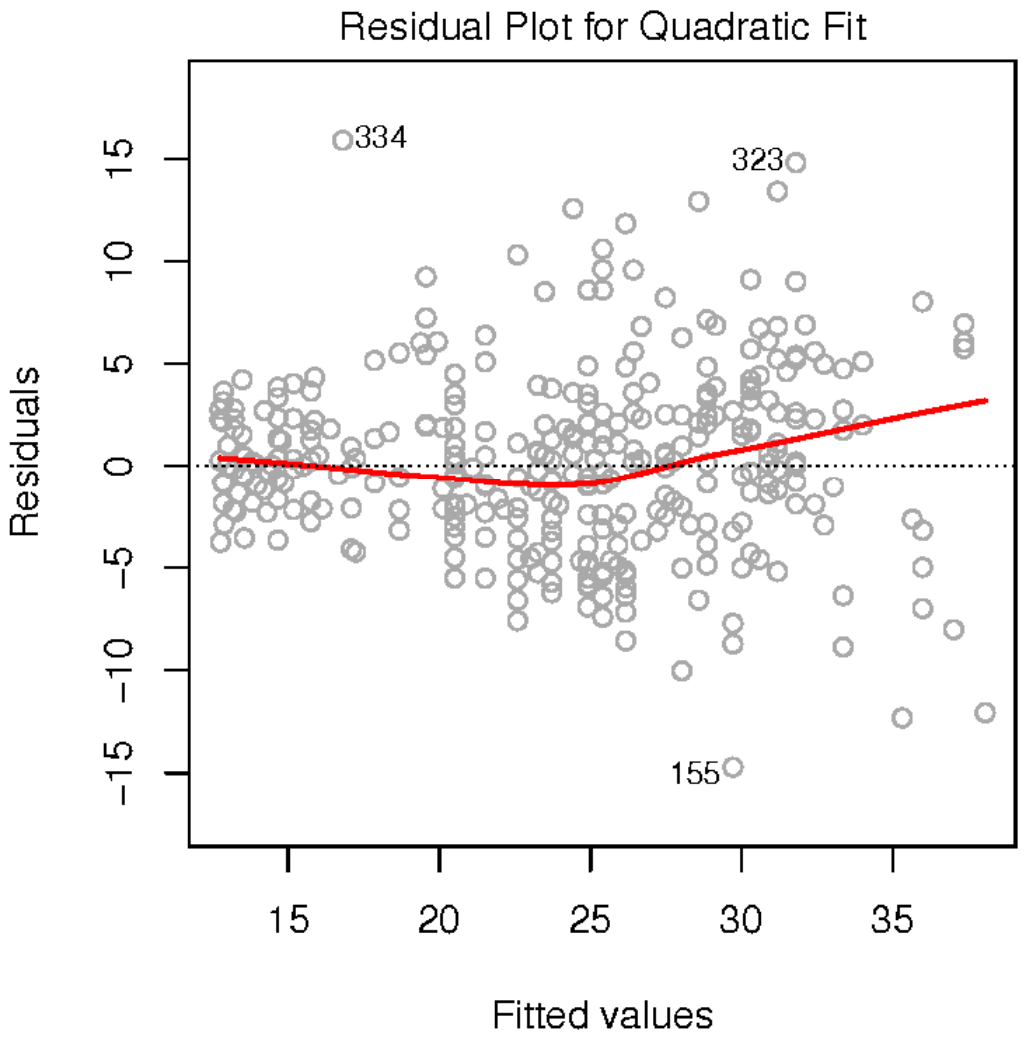
- 비선형성
- 상관된 오차
- 이분산성
- 이상점
- 지렛점
- 공선성

비선형성

- 잔차 플롯(residual plots) 이용: 잔차($e_i = y_i - \hat{y}_i$) 대 x_i 그림 (단순회귀모형) 혹은 잔차 대 $\hat{y}_i$ 그림 (다중회귀모형)
- 눈에 띄는 경향이 없으면 선형성 만족!
- 그러나 어떤 경향을 보이면 예측변수를 비선형(non-linear) 변환 (가령, $\log X, \sqrt{X}, X^2$ 등) 시도
- 예: 자동차 자료
  - 단순선형모형 vs. 2차 다항회귀모형

그림3-9: 자동차 자료. 왼쪽(단순회귀모형), 오른쪽(2차 선형회귀모형)

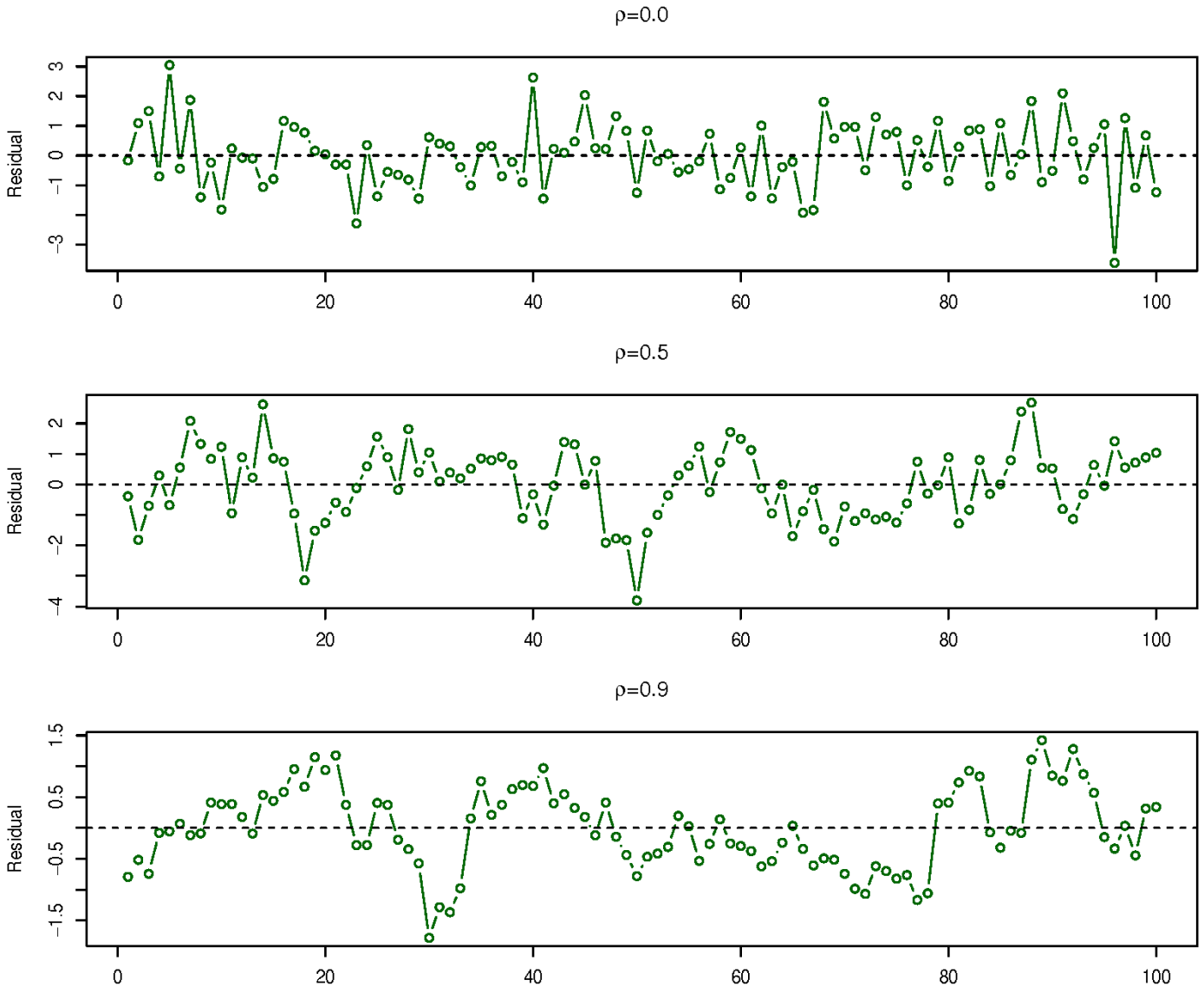




상관된 오차

- 오차들($\epsilon_1, \epsilon_2, \ldots, \epsilon_n$)이 독립이 아니고 서로 상관되면(correlated) 무슨 일이 생기나?
  - 추정량의 SE가 과소추정(underestimate)됨!
  - 신뢰구간의 폭이 좁아져서 명목값인 0.95에 미치지 못함
  - 검정통계량의 절대값이 커져 p-값이 작아져 H_0를 기각하는 잘못된 결정을 하게 됨
- 시계열(time series) 자료에서 흔히 발생 혹은 가족 구성원에 대한 자료, 동일 환경에 노출되어 있는 개체 자료 등
- 시간에 대한 잔차 플롯을 그려 판단

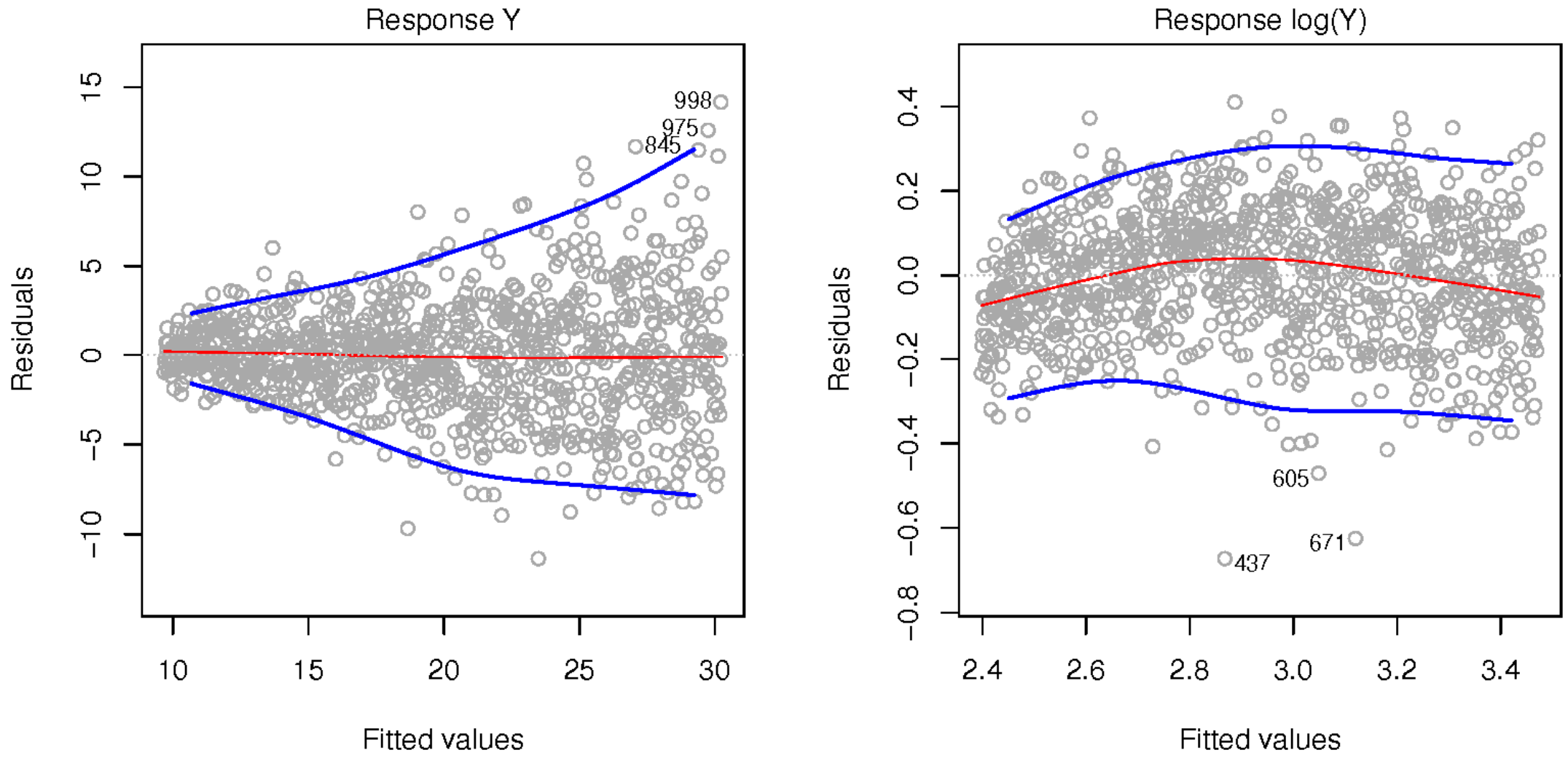
그림3-10: 가상 자 료 (위 쪽 - 무상관, 중앙- 보통, 아래쪽- 강한 양의 상관)



이분산성

- 반응변수 값이 커질수록 오차의 분산도 커지는 경우가 발생 ⇒ 이분산성(heteroscedacity)
- 잔차 플롯 형태가 굴뚝 모양(funeral shape)!
- 해법: 반응변수를 $\log Y$, $\sqrt{Y}$와 같은 오목(concave)함수로 변환!
- 예: i-번째 반응변수가 n_i개 개체들의 평균이라면 분산이 $\frac{\sigma^2}{n_i}$이므로 분산이 다름
  - 가중(weighted)최소제곱법 이용!
  - 가중값은 $w_i = n_i$

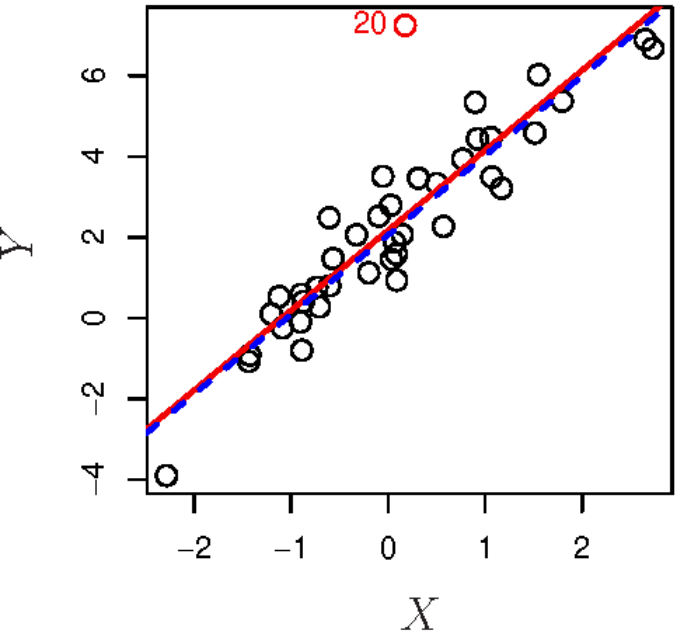
그림3-11: 왼쪽(원자료), 오른쪽(log Y 변환); 빨간색(잔차를 평활)

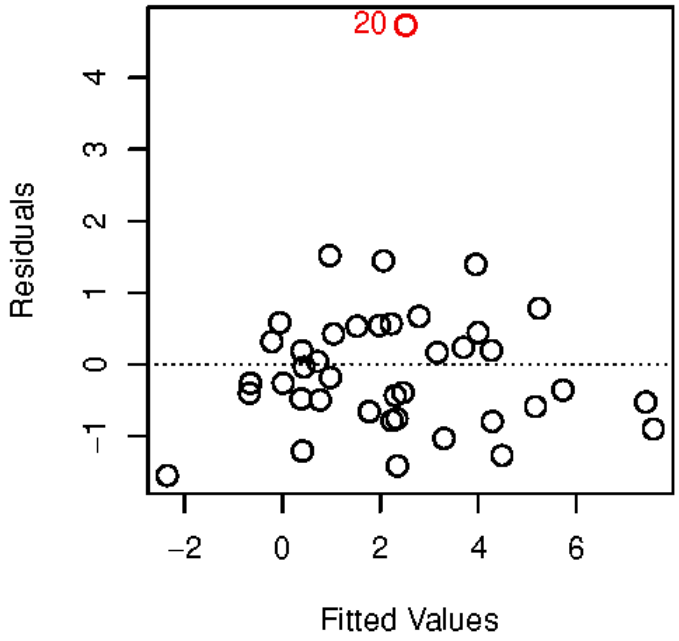


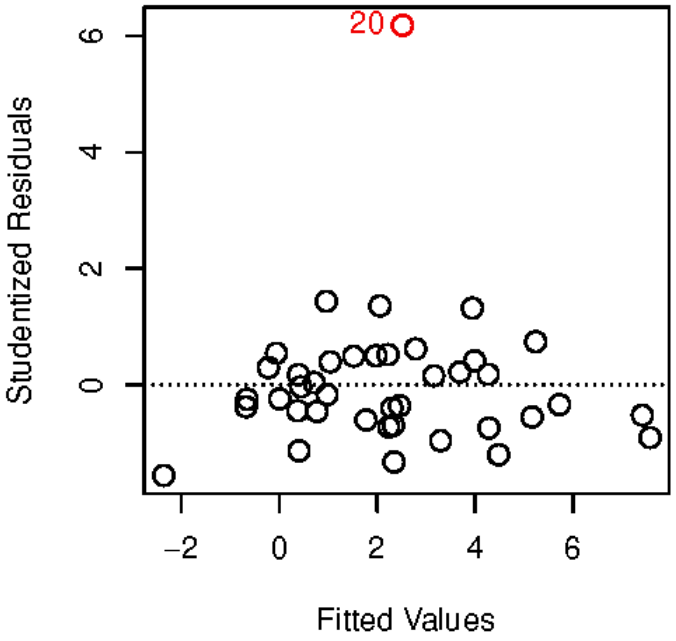
이상점(outliers)

- y_i가 $\hat{y}_i$로부터 많이 떨어져 있을 때
- 최소제곱추정량에는 거의 영향을 미치지 않을 수도! 그러나 RSE는 커질 수도! R^2는 작아질 수도!
- 잔차 플롯보다 표준화잔차 플롯을 그려 $(-2, 2)$ 범위 밖에 있는 것은 이상점으로 간주!
- 단순히 제거하기보다 모형에 부족(deficiency)한 것(예를 들어 모형에 포함하지 않은 예측변수 등)은 없는지 검토 필요

그림3-12: 빨간색(20 포함), 파란색(20 미포함)



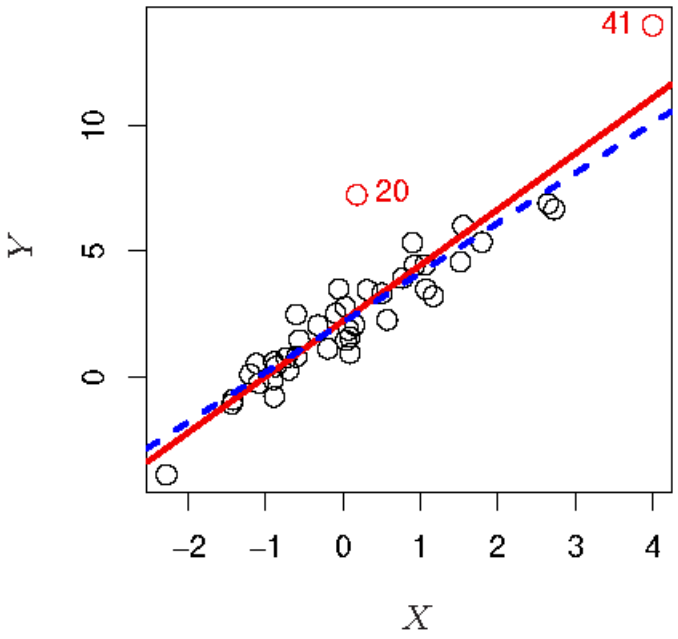


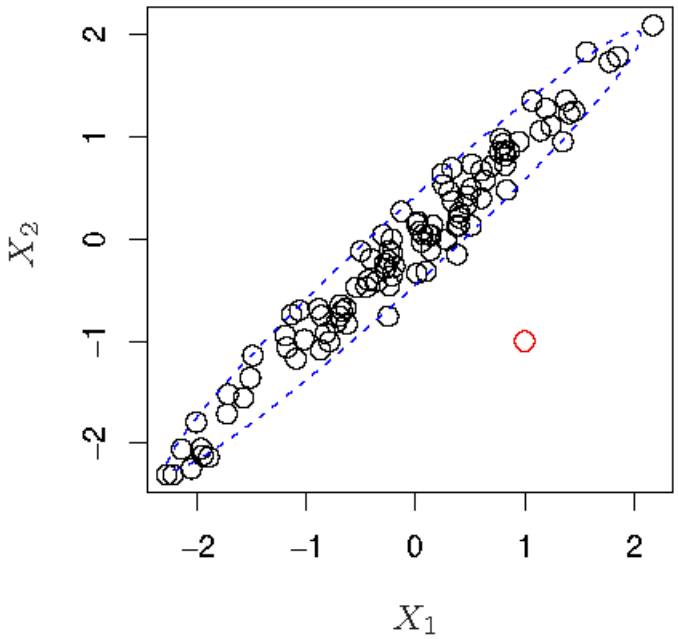


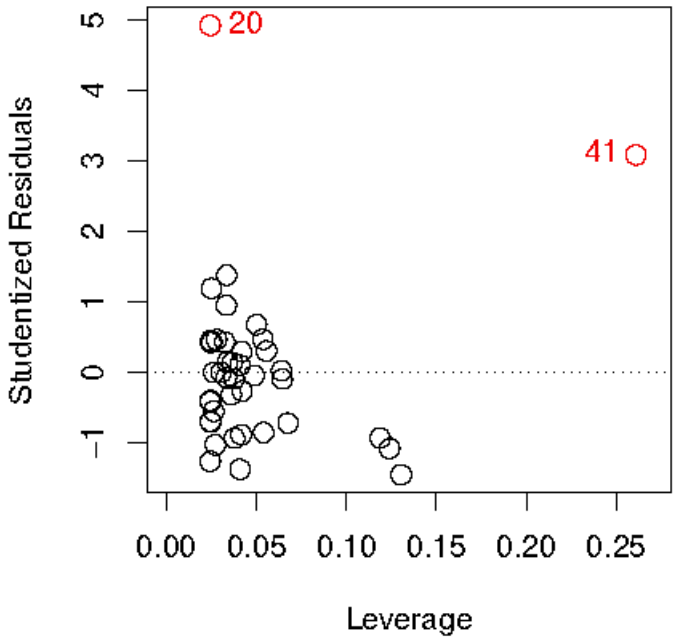
지렛점(high leverage points)

- 유별난 x_i를 가진 개체
- 최소제곱추정량에 영향을 줄 수 있음
- 예측변수가 1개이면 쉽게 지렛점을 찾을 수 있지만, 2개 이상이면 레버리지(leverage) 통계량을 계산
  - 예측변수가 1개일 때, 레버리지 통계량: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \in \left(\frac{1}{n}, 1\right)$
- 평균 레버리지가 $\frac{p+1}{n}$이기 때문에 이 값보다 큰 값을 가진 개체는 지렛점으로 의심해야!

그림3-13: 왼쪽(20-이상점, 41-지렛점, 이상점), 중앙(결합 지렛점)



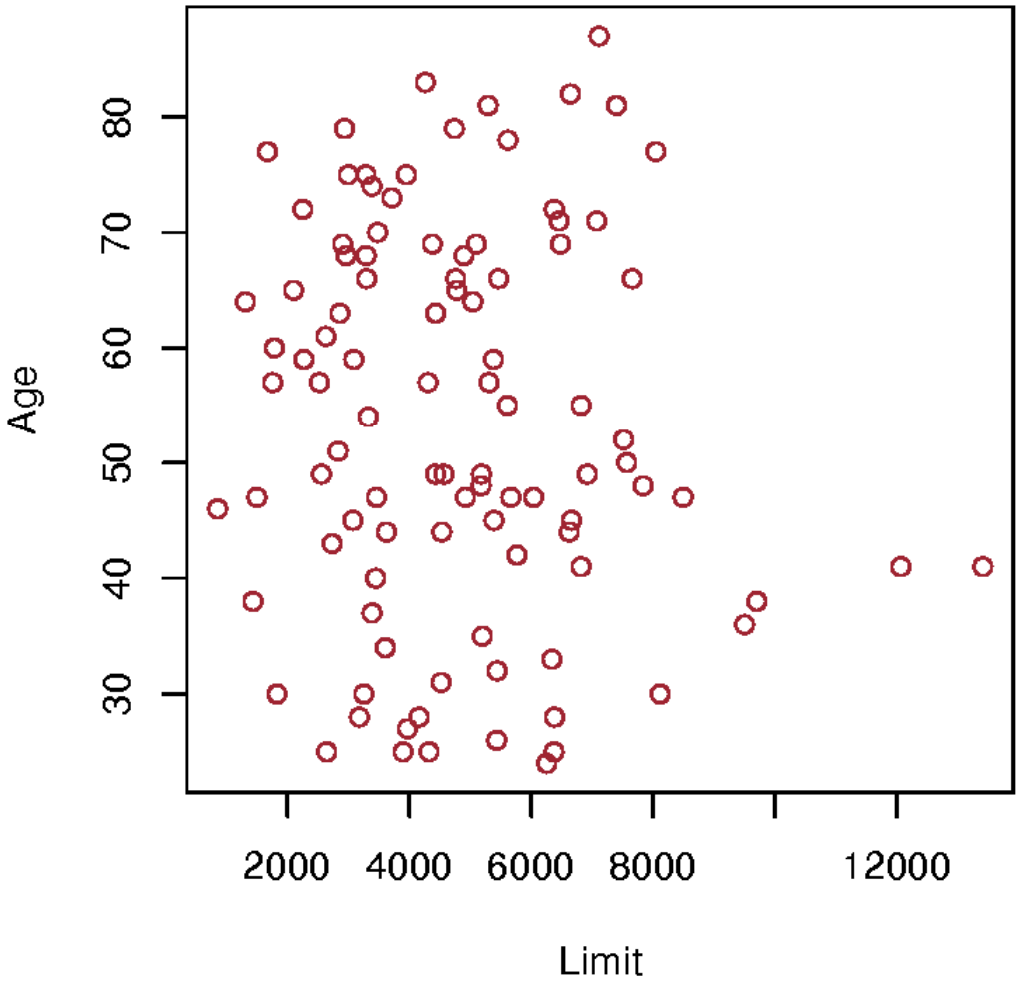




공선성(collinearity)

- 둘 혹은 셋 이상의 예측변수들이 밀접하게 연관되었을 때
- 예: 신용 자료
  - 예측변수: 한도와 연령 ⇒ 무상관, RSS의 등고선이 동그란 모양
  - 예측변수: 한도와 신용등급 ⇒ 매우 높은 상관(collinear라고 부름), RSS의 등고선이 좁은 협곡(narrow valley) 모양 즉, LSE로 가능한 값의 범위가 넓음. 자료값이 일부만 바뀌더라도 LSE가 쉽게 변함

그림3-14: 왼쪽(무상관), 오른쪽(강한 양의 상관)



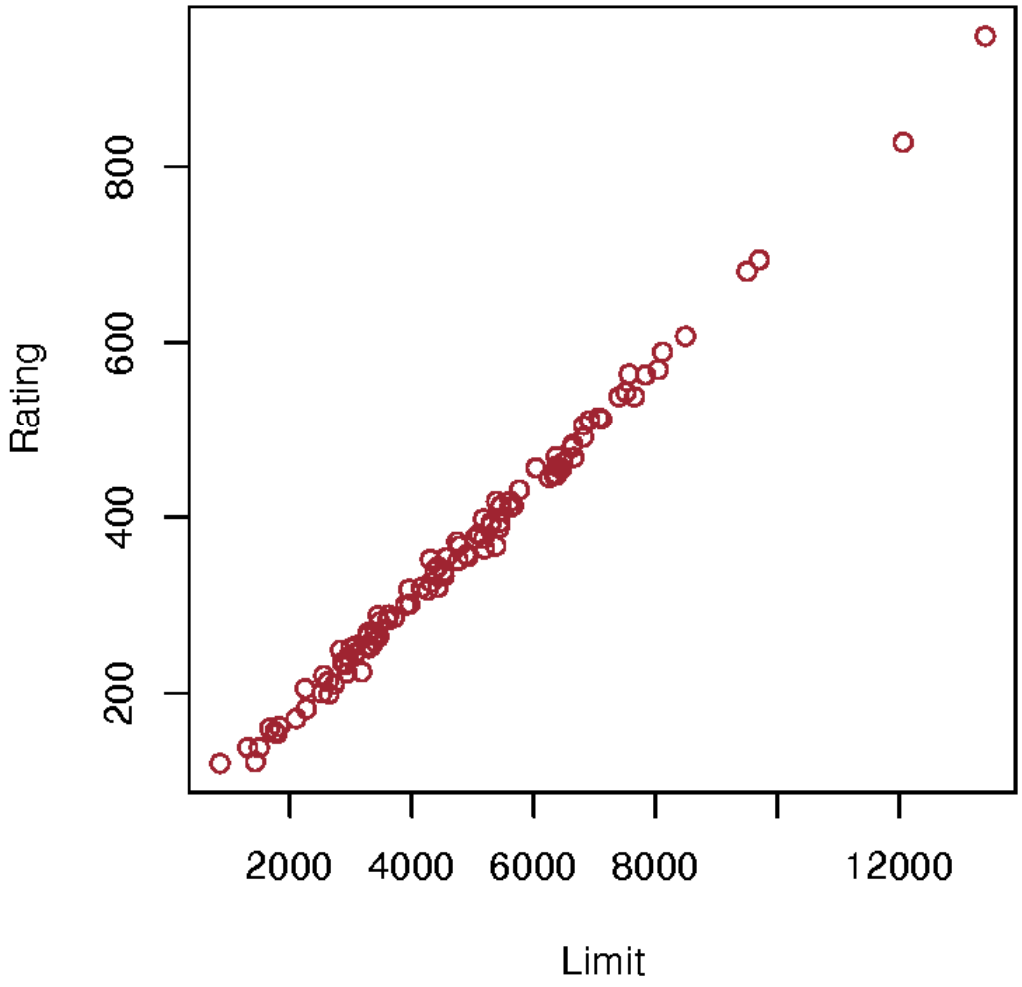
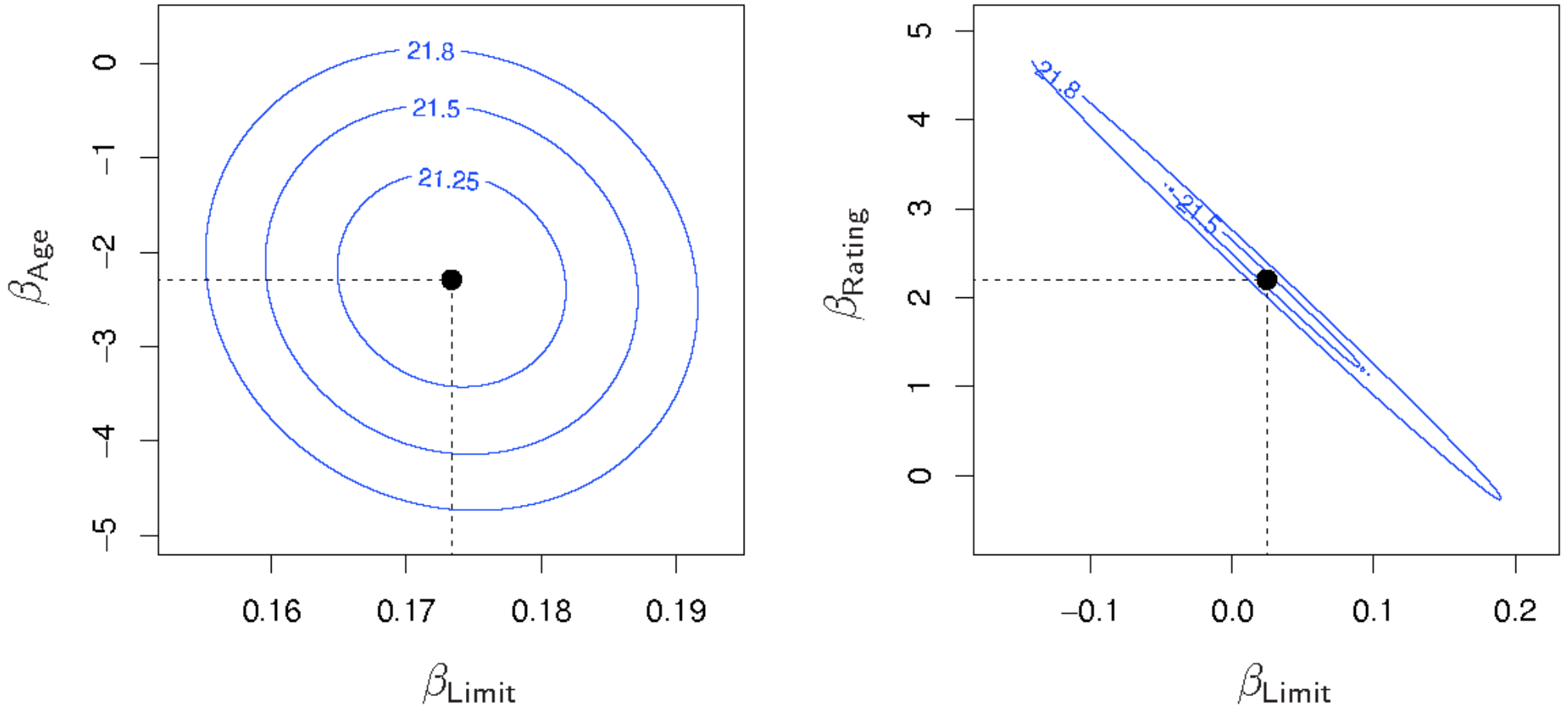


그림3-15: 왼쪽(무상관), 오른쪽(강한 양의 상관)



공선성이 존재하면 …

- 회귀계수 추정량의 정확도(accuracy)가 떨어져 추정량의 표준오차가 증가
- t-통계량 값을 작게 만들어 $H_0: \beta_j = 0$을 기각하지 못하게 될 수도!

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

Model 2에서 limit의 표준오차는 매우 큼(Model 1보다 12배). 따라서 t-통계량 값은 작아지고 p-값은 커짐.

공선성을 찾는 방법

- 상관계수행렬 이용: 상관계수의 절대값이 큰 쌍은 의심
- 다중공선성(multicollinearity)이 존재할 때 즉 셋 이상의 예측변수들이 상관되었을 때는 상관계수행렬로 탐지가 어려움
- 분산팽창계수(variance inflation factor, VIF) 이용!

분산팽창계수

- 정의: $\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}} \in (1, \infty)$
- $R^2_{X_j|X_{-j}}$: X_j를 반응변수로 하고 나머지 예측변수를 예측변수로 하는 회귀모형의 R^2
- 10 이상인 예측변수는 다중공선성이 존재한다고 의심

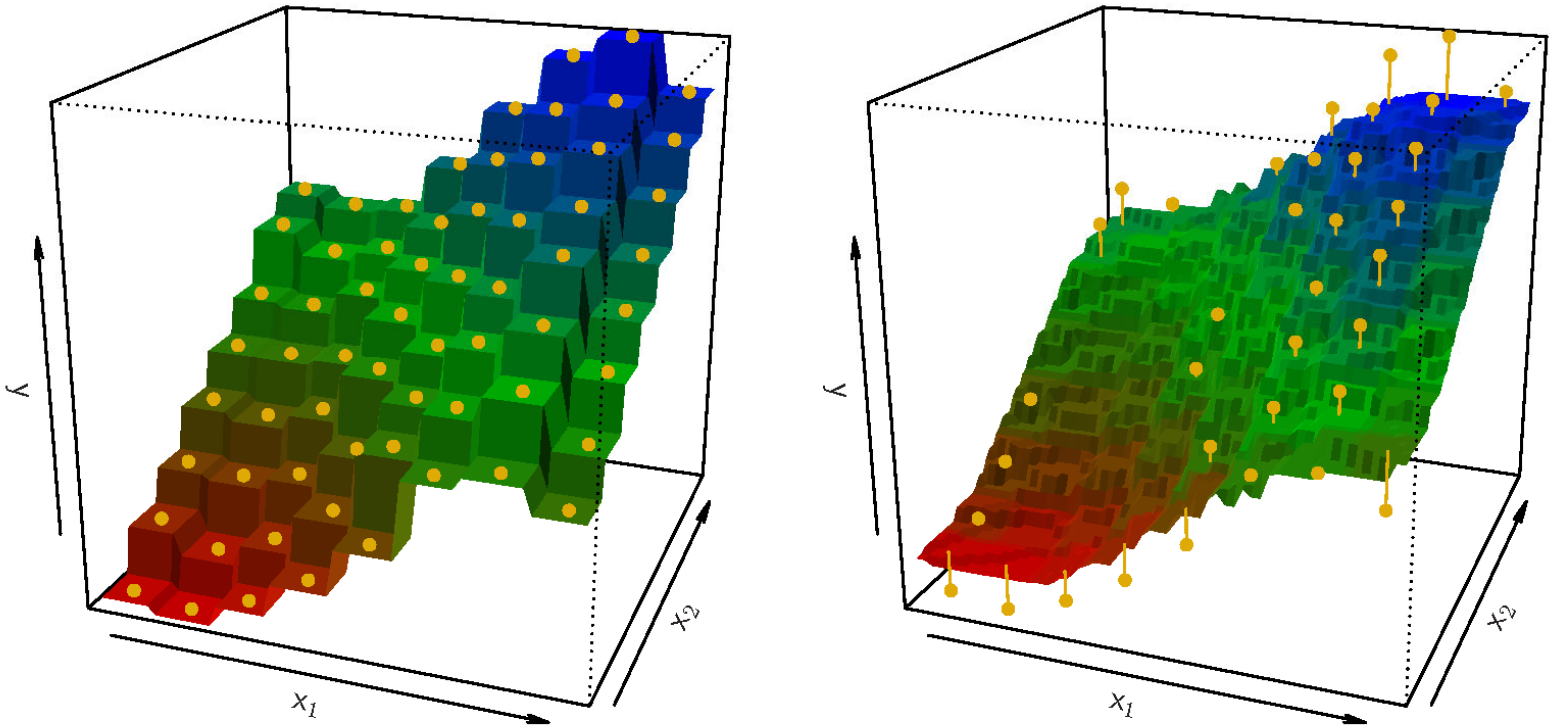
공선성을 해결하는 방법

- 다중공선성이 있는 변수들 중에서 한 개를 모형에서 제거
- 다중공선성이 있는 변수들을 결합하여 새로운 한 변수로 변환

KNN과 비교

- 미지의 $f(X)$에 대해 어떤 함수 꼴을 가정하지 않기 때문에 비모수적이고 선형모형보다 더 유연!
- KNN 예측값: $\hat{f}(x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} y_i$
  - $\mathcal{N}_0$: x_0 근방에 있는 훈련세트 개체의 집합
- K의 값이 작으면 편의는 줄어들지만 분산은 늘어나고, K의 값이 크면 분산은 줄어들지만 편의 늘어나고 미지의 $f(X)$를 가릴(masking) 수도! (그림3.16 참조)
- 최적 K를 구하는 문제 중요! 5장에서 다룸!

그림3-16: 왼쪽(K=1), 오른쪽(K=9)



KNN과 비교

- 언제 모수적 방법이 더 우수한가? 가정한 모형이 자료에 잘 맞을 때
- 선형적이면 선형모형이 KNN보다 우수하고, 비선형이면 그 반대! (그림3.18, 3-19 참조)
- 전체적으로 KNN은 선형모형보다 우수하다고 할 수 있지만 예측변수의 개수가 늘어날수록 선형모형이 우수!
  - 예: $n = 100, p = 20$일 때 최근접 개체가 거의 없을 수도! ⇒ KNN 적합 결과 좋지 않음. 차원(p)의 저주(curse of dimensionality)라 부름!
- 예측변수의 개수가 적을 때도 해석에 초점을 맞추면 선형모형 선호!

그림3-17: 왼쪽(K=1), 오른쪽(K=9). 검은색(True 회귀직선)

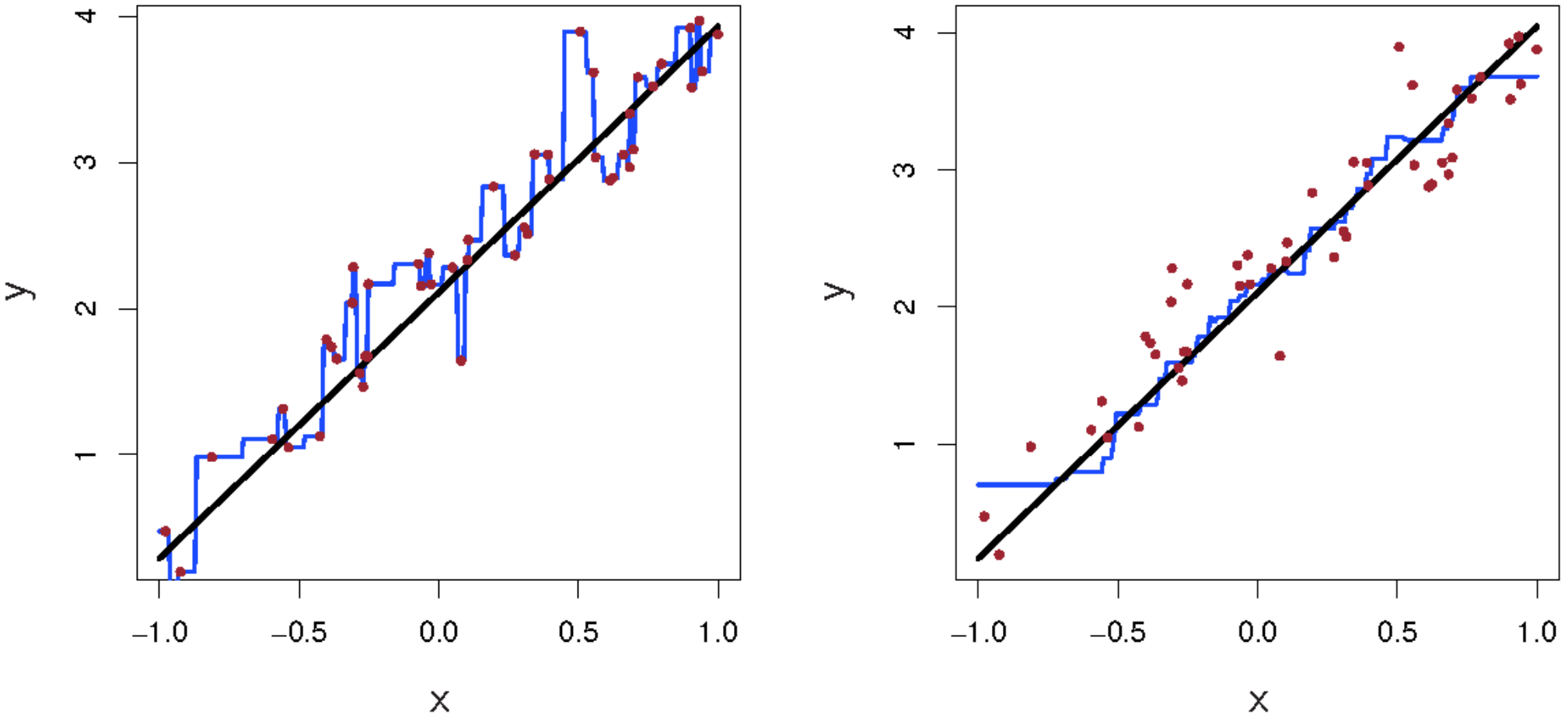


그림3-18: 파란색(LS 회귀직선). 녹색(KNN MSE). 점선(LS MSE)

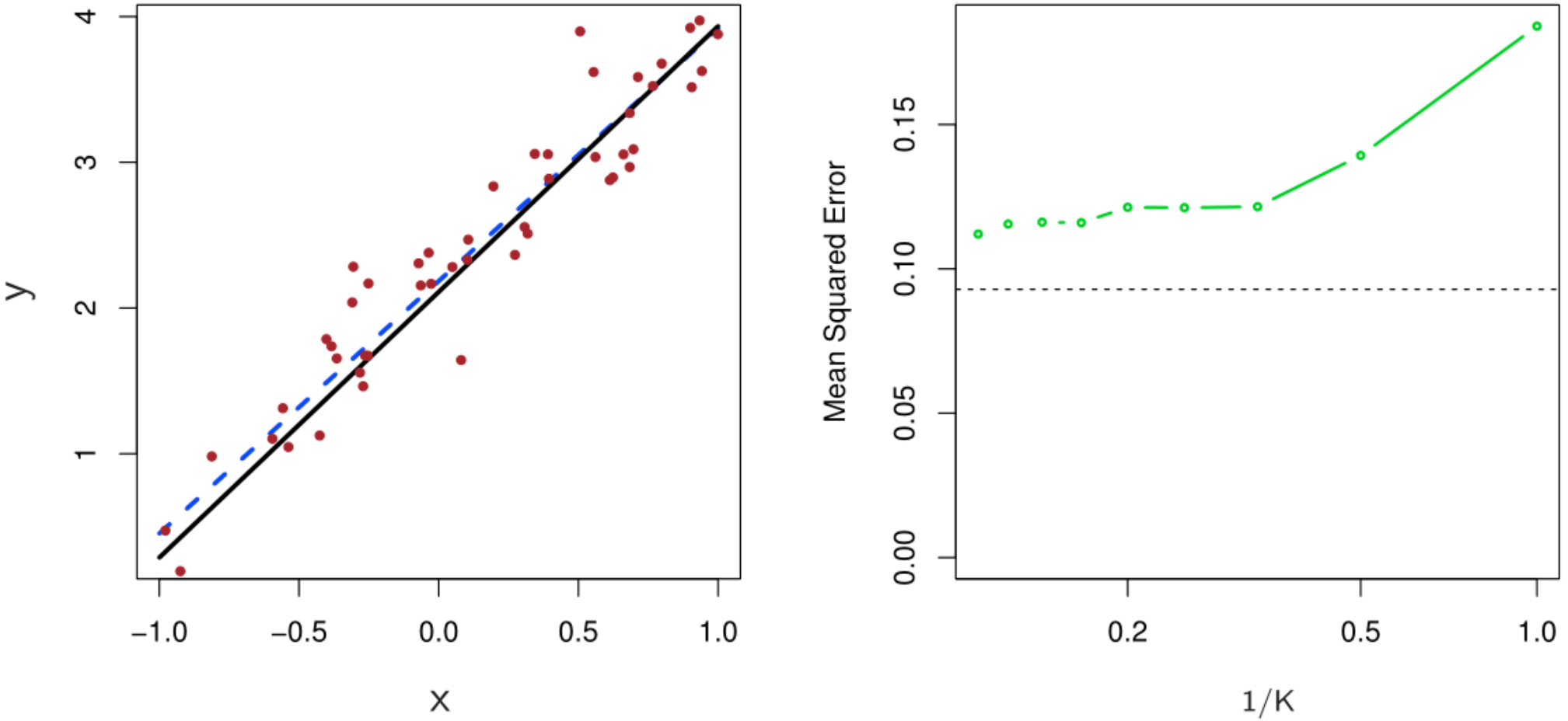


그림 3-19: 위쪽(다소 비선형), 아래쪽(비선형). 검은색(True 회귀선), 파란색 (K=1), 빨간색 (K=9). 녹색 (KNN MSE). 점선(LS MSE)

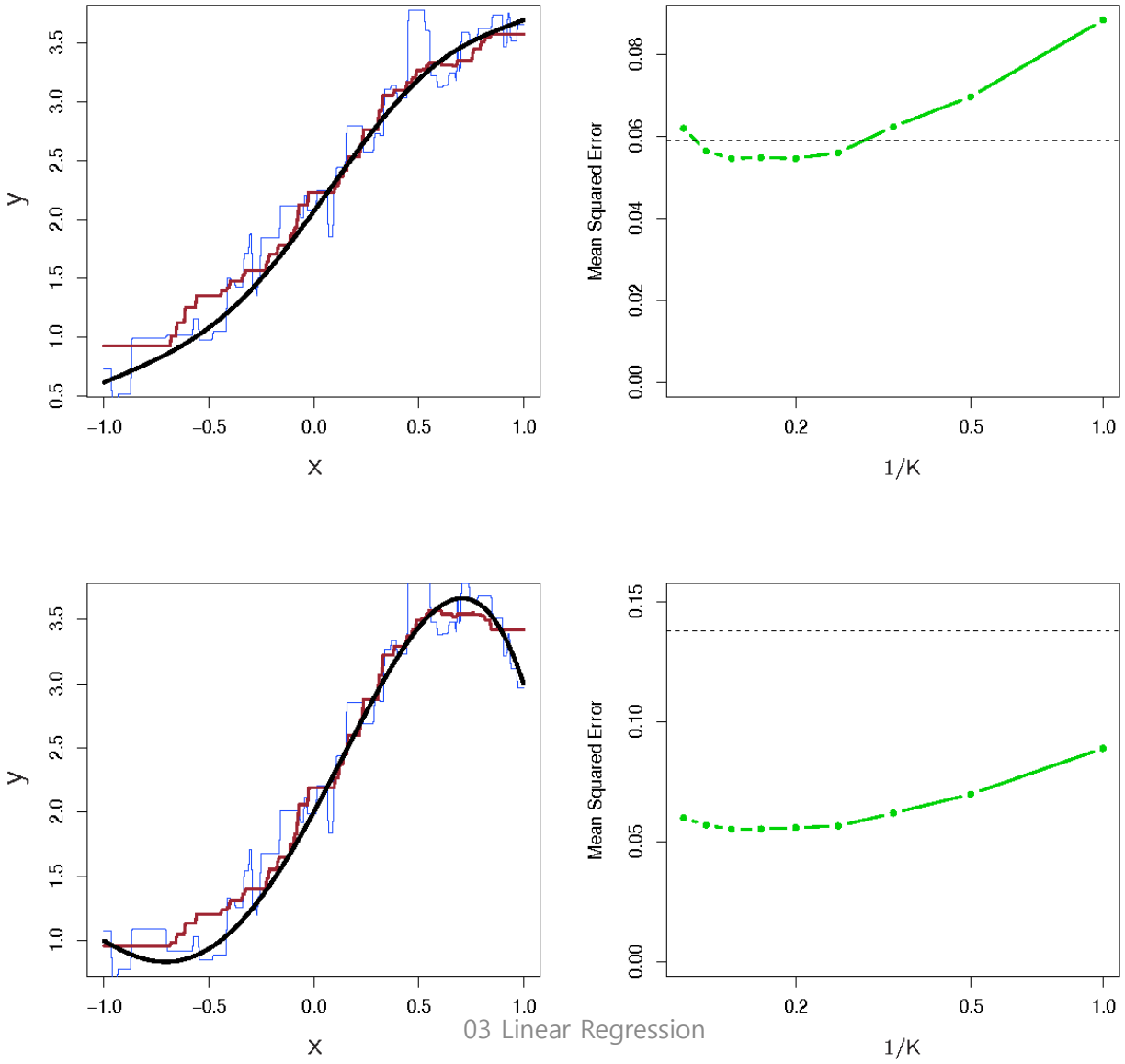
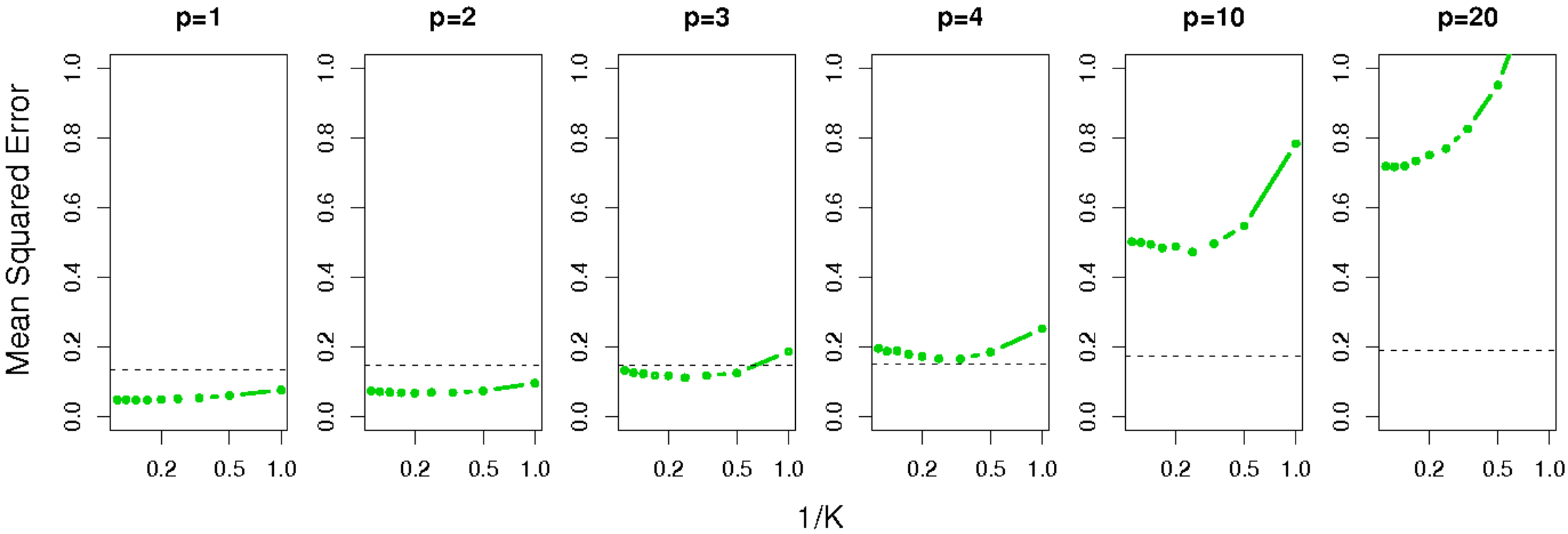


그림3-20: 첫 번째 예측변수의 True 회귀선(그림3-19 아래쪽), 나머지 예측변수는 Y와 관계 없음. 녹색(KNN MSE). 점선(LS MSE)



과제(4월25일 마감)

- 연습문제3장: 3, 4, 11, 13, 15

Thank you!

Move on to 04 Classification

04 Classification

J Kim

2021.4

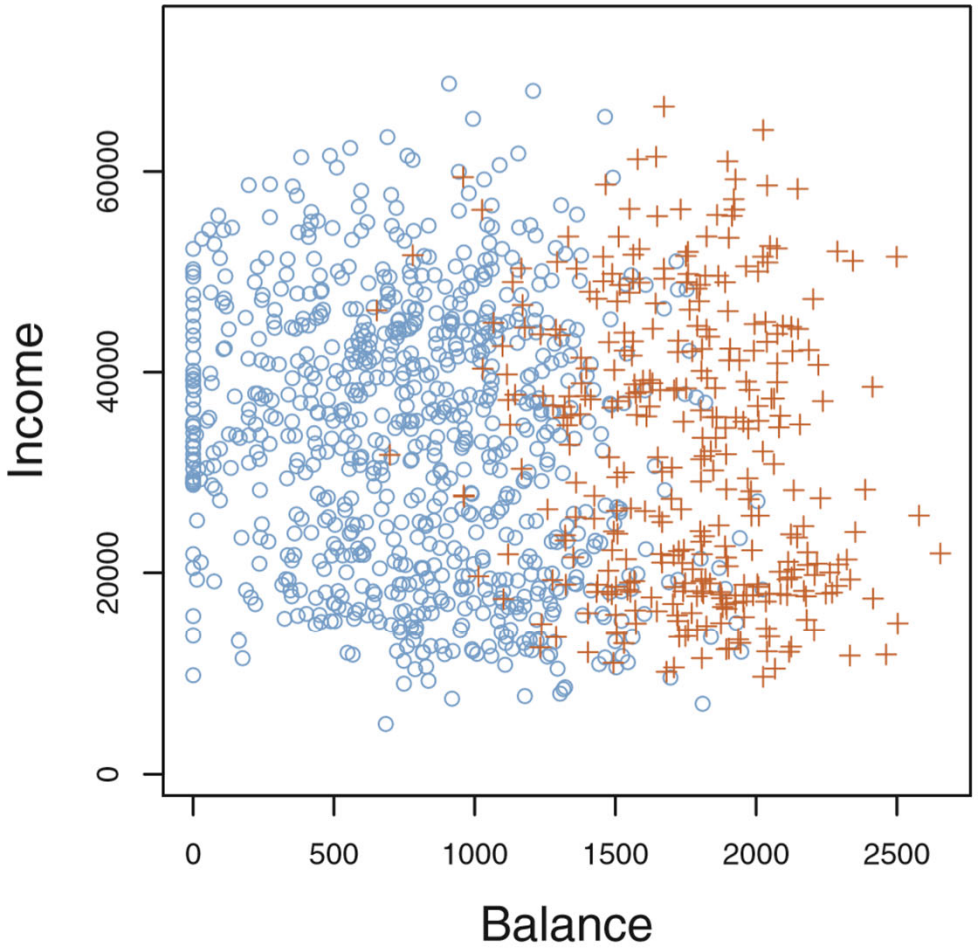
Outline

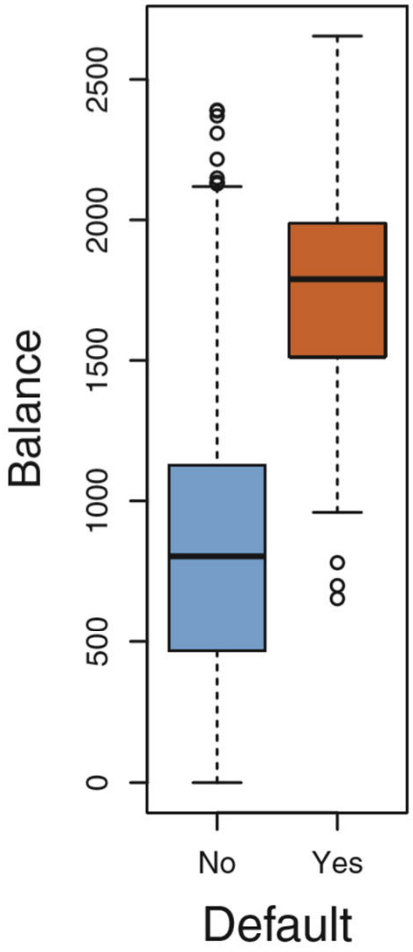
로지스틱 회귀모형
선형판별분석
이차판별분석
KNN과 비교

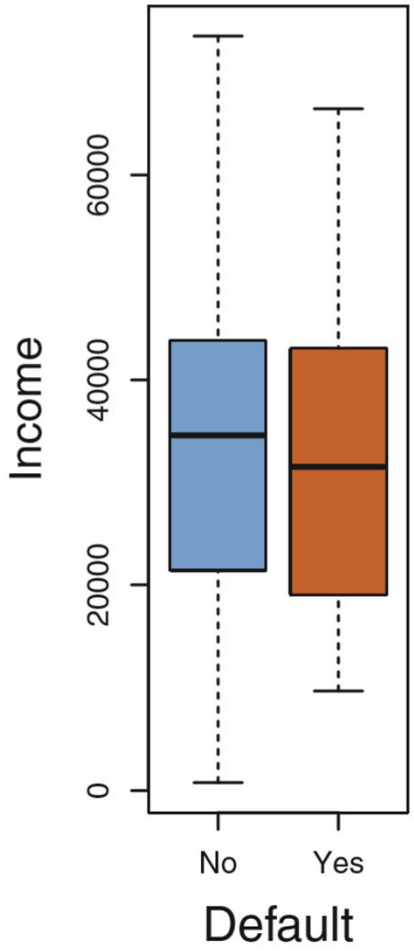
Default(채무 불이행) 자료

- A data frame with 10000 observations on 4 variables
- default A factor with levels No and Yes indicating whether the customer defaulted on their debt
- student A factor with levels No and Yes
- balance The average balance that the customer has remaining on their credit card after making their monthly payment
- income Income of customer

그림4.1 채무 불이행 자료. 파란색(채무 불이행 아님) , 오렌지색(채무 불이행)



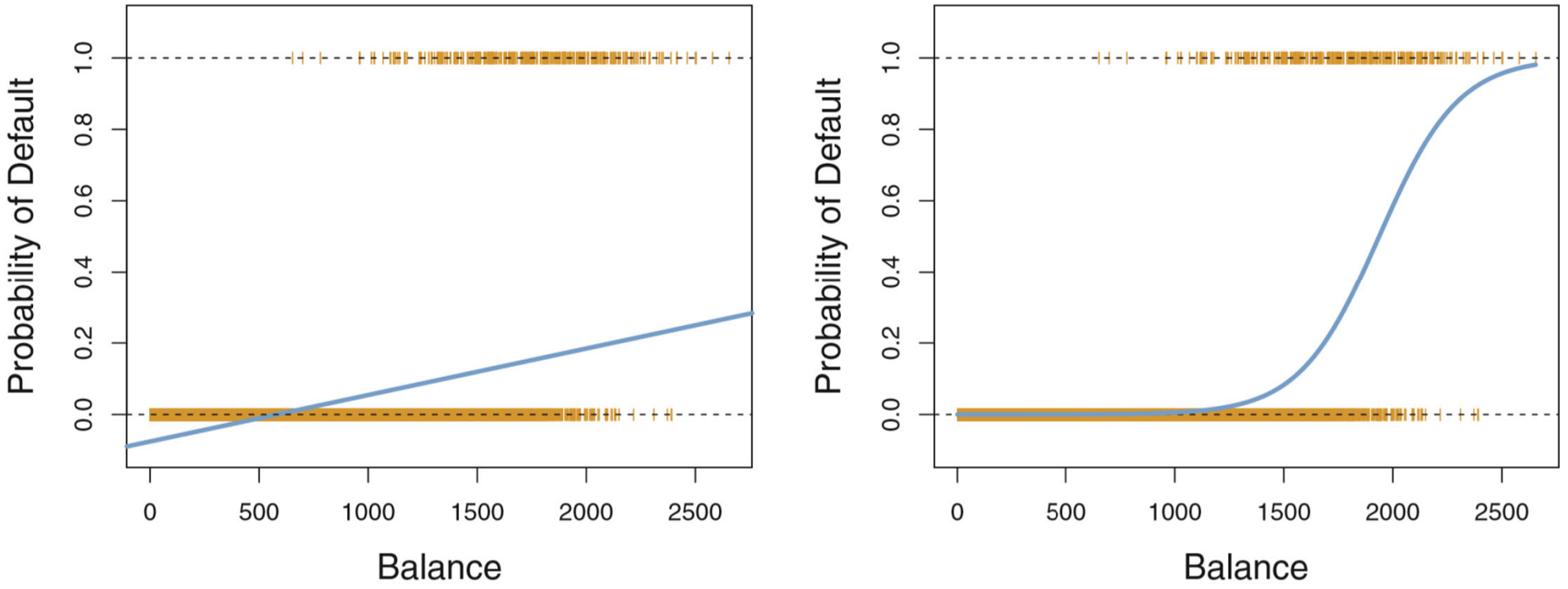




왜 선형회귀모형(LR)은 안 되는가?

- Y가 binary(yes/no)일 때 LR을 적용하면 $\hat{Y} = X\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$은 $P(Y = 1|X)$에 대한 예측값을 의미하므로 "$\hat{Y} > 0.5$"이면 "yes"로 분류하고 그렇지 않으면 "no"로 분류. 다만 $\hat{Y} \in [0,1]$을 만족해야!
- 하지만 채무 불이행 자료에서 balance가 578보다 작으면 $\hat{Y} < 0$! balance가 8278보다 크면 $\hat{Y} > 1$!
- Y의 범주가 3개 이상일 때 LR은 분류 문제에 적용할 수 없음!

그림4.2 채무 불이행 자료. 왼쪽(선형회귀모형으로 적합), 오른쪽(로지스틱 회귀모형으로 적합)



로지스틱 회귀모형(logistic regression)

- Y를 예측하기보다 Y가 각 범주에 속할 확률을 예측! ⇒ 분류 문제 활용 가능
- $p(X) = P(Y = 1|X)$와 X를 어떻게 연결할 것인가?
- 로지스틱 함수: $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \in (0,1)$
- β_1이 양수이면 X가 증가함에 따라 $p(X)$가 증가하고, β_1이 음수이면 X가 증가함에 따라 $p(X)$가 감소함
- "S"자 형태!

로지스틱 회귀모형

- 오즈(odds): $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \in (0, \infty)$
- 로그 오즈(log odds) 혹은 로짓(logit): $\log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X \Rightarrow$ 로지스틱 회귀모형(logistic regression)
- X를 한 단위(one unit) 증가 시키면 로그 오즈는 β_1만큼 변하고, 오즈는 e^{β_1}배 됨!
- $\text{logit}(X = x + 1) - \text{logit}(X = x) = \beta_0 + \beta_1(x + 1) - \beta_0 - \beta_1 x = \beta_1$
- $\frac{\text{odds}(X=x+1)}{\text{odds}(X=x)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$

회귀계수 추정

- 최대우도(maximum likelihood) 원리
- 우도 함수(likelihood function):

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

$$= \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$$

- 최대우도추정량(MLE): $l(\beta_0, \beta_1)$ (혹은 $\log l(\beta_0, \beta_1)$)를 최대로 하는 추정량 $\Rightarrow \hat{\beta}_0, \hat{\beta}_1$

회귀계수 추정

- Remarks 1: 표4.1에서 z-통계량은 가설 $H_0: \beta_j = 0\ (j = 0,1)$을 검정하기 위한 통계량. LR에서 t-통계량과 유사
- Remarks 2: 범주형 예측변수는 더미변수를 이용하여!

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

TABLE 4.1. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

신용카드 빚이 1단위(천불) 늘어나면 채무 불이행할 로그오즈는 0.0055 만큼 늘어나고, 채무 불이행할 오즈는 $e^{0.0055} = 1.006$배 됨

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

TABLE 4.2. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using student status. Student status is encoded as a dummy variable, with a value of* 1 *for a student and a value of* 0 *for a non-student, and represented by the variable* `student[Yes]` *in the table.*

학생이 아닌 사람보다 학생은 채무 불이행할 로그오즈는 0.40490055 만큼 늘어나고, 채무 불이행할 오즈는 $e^{0.4049} = 1.5$배 됨

예측

- $\hat{p}(X) = \hat{P}(Y = 1|X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$: 예측 확률
- 만일 "$\hat{p}(X) > 0.5$" 이면 범주 "1"로 분류하고 그렇지 않으면 범주 "0"으로 분류

다중 로지스틱 회귀모형

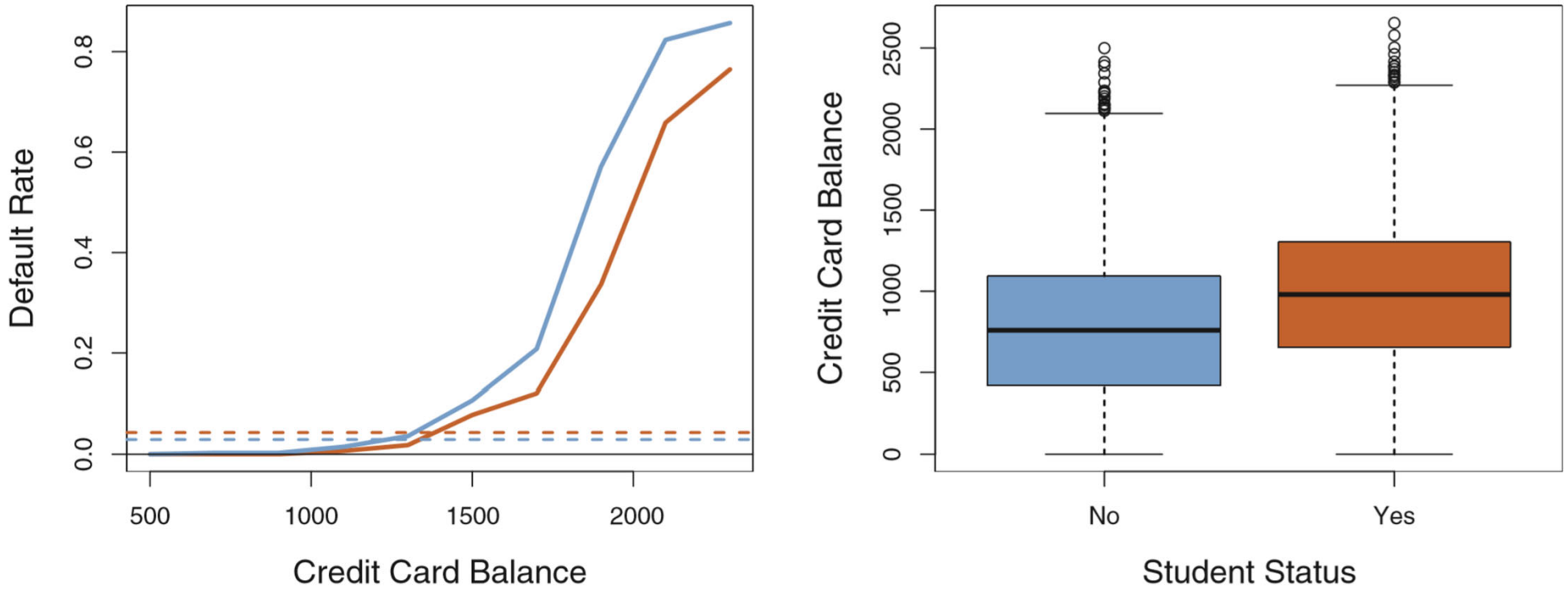
- $p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$: 로지스틱 함수
- $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$: 오즈
- $\log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$: 다중 로지스틱 회귀모형
- MLEs: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance, income, *and student status. Student status is encoded as a dummy variable* student[Yes], *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

Table 4.2에서는 채무 불이행할 확률이 학생이 높았지만 Table 4.3에서는 정반대임

그림4.3 채무 불이행 자료. 파란색(학생 아님) , 오렌지색(학생). 실선(신용카드 빚에 따라 채무 불이행 비율), 점선(전체 채무 불이행 비율)



점선은 Table 4.2의 결과이며 실선은 Table 4.3의 결과(?). 신용카드 빚을 고려하지 않으면 학생이 채무 불이행 가능성이 높지만 신용카드 빚을 고려하면 정반대임. 이는 학생 유무와 신용카드 빚이 연관되어서. 즉 두 변수가 교락(confounding)되어 있음

범주가 3개 이상인 로지스틱 회귀모형

- 기저(baseline) 범주와 나머지 범주에 대해 로지스틱 회귀모형을 구축. 모형 개수는 범주수보다 1개 적음
- 그러나 선형판별분석을 권장!

선형판별분석(linear discriminant analysis, LDA): 왜 LDA가 필요한가?

- 범주에 따라 예측변수의 값이 분리되어 있으면 로지스틱 회귀모형의 추정량은 불안정
- n이 작고 예측변수 X가 근사적으로 정규분포를 따르면 LDA가 로지스틱 회귀모형보다 더 안정적
- 반응변수의 범주가 3개 이상일 때 다루기 쉬움!

선형판별분석

- $K(\geq 2)$: 범주수
- $\pi_k(k = 1,2, \dots, K)$: 사전(prior) 혹은 통합(overall) 확률. 즉 무작위로 선택된 한 개체가 k-번째 범주에서 나올 확률
- $f_k(x) = P(X = x|Y = k)$: k-번째 범주에 속한 개체의 예측변수 X의 확률밀도함수
- $p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$: 사후(posterior) 확률. 즉 $X = x$를 가진 한 개체가 k-번째 범주에 속할 확률

LDA 분류기

- 만일 "$k^* = \text{argmax}_{1 \le k \le K} p_k(x)$" 이면 $X = x$를 가진 개체를 "$k^*(k^* = 1,2,\ldots,K)$" 범주로 분류 ⇒ 베이즈 분류기
- $\hat{p}_k(x)$의 추정
  - $\hat{\pi}_k$: 훈련세트의 범주별 표본비율
  - $\hat{f}_k(x)$: 어떻게? 분포를 가정하여!
- 만일 "$\hat{k}^* = \text{argmax}_{1 \le k \le K} \hat{p}_k(x)$" 이면 $X = x$를 가진 개체를 "$\hat{k}^*(k^* = 1,2,\ldots,K)$" 범주로 분류 ⇒ LDA 분류기

예측변수가 1개일 때

- 가정 1 (정규성): $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}, k = 1,2,\dots,K$
- 가정 2 (등분산성): $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2 \Rightarrow$ 강한 가정!
- 베이즈 분류기
  - "$\delta_k(x) = \frac{\mu_k}{\sigma^2} \times x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \ (k = 1,2,\dots,K)$"를 가장 크게 하는 범주로 분류 WHY?
- 베이즈 결정 경계: $\{x: \delta_k(x) = \delta_l(x), k \neq l\}$

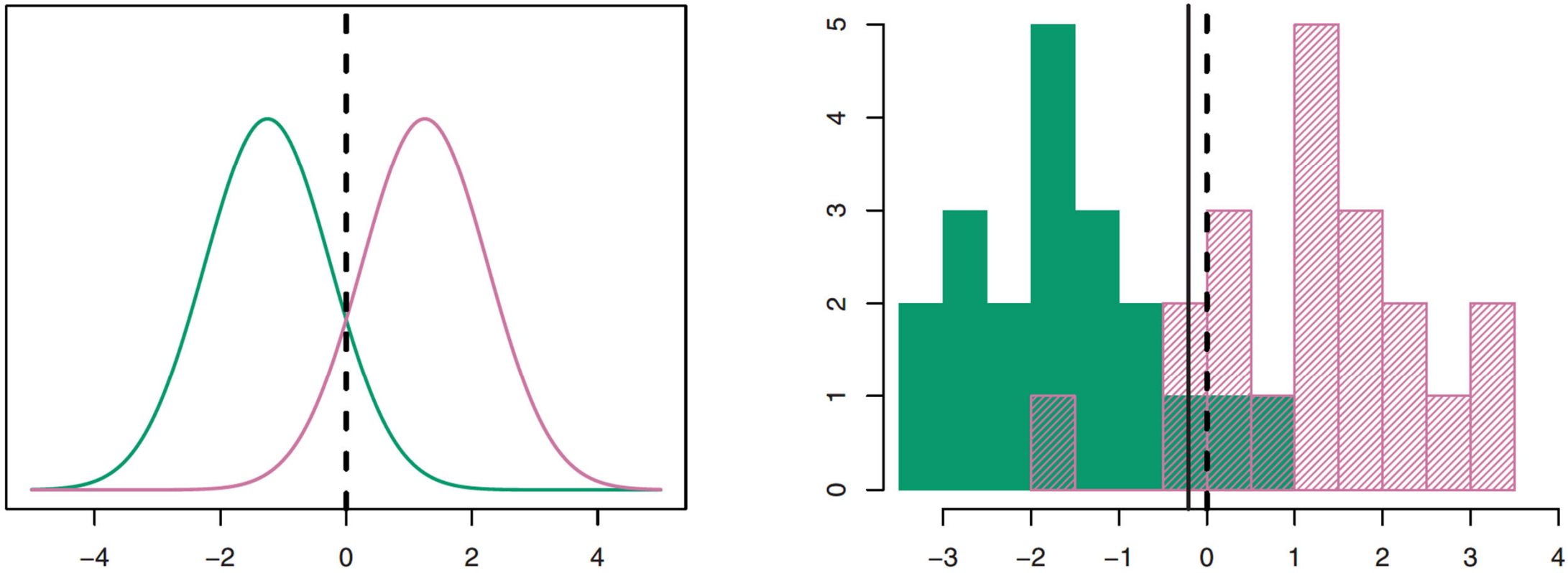
예측변수가 1개일 때

- LDA 분류기
  - "$\hat{\delta}_k(x) = \frac{\hat{\mu}_k}{\hat{\sigma}^2} \times x - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k \ (k = 1,2, \dots, K)$"를 가장 크게 하는 범주로 분류
  - $\hat{\pi}_k = \frac{n_k}{n}$, $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$, $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$
- LDA 결정 경계: $\{x: \hat{\delta}_k(x) = \hat{\delta}_l(x), k \neq l\}$

예측변수가 1개일 때

- 예: $K = 2, \pi_1 = \pi_2, f_1(x) \sim N(-1.25,1), f_2(x) \sim N(1.25,1)$ 일 때,
- "$x < \frac{\mu_1+\mu_2}{2} = 0$"이면 범주 1로 분류하고 그렇지 않으면 범주 2로 분류 WHY?
  - 베이즈 결정경계: $\frac{\mu_1+\mu_2}{2} = 0$, 베이즈 오류율: 10.6% WHY?
- "$x < \frac{\bar{x}_1+\bar{x}_2}{2}$"이면 범주 1로 분류하고 그렇지 않으면 범주 2로 분류
  - LDA 결정경계: $\frac{\bar{x}_1+\bar{x}_2}{2}$, 테스트 오류율: 11.1%

그림4.4 녹색(평균=-1.25, 표준편차=1인 정규분포), 분홍색(평균=1.25, 표준편차=1인 정규분포) . 20개 랜덤표본의 히스토그램. 점선(베이즈 결정 경계 값), 실선(LDA 결정경계 값)



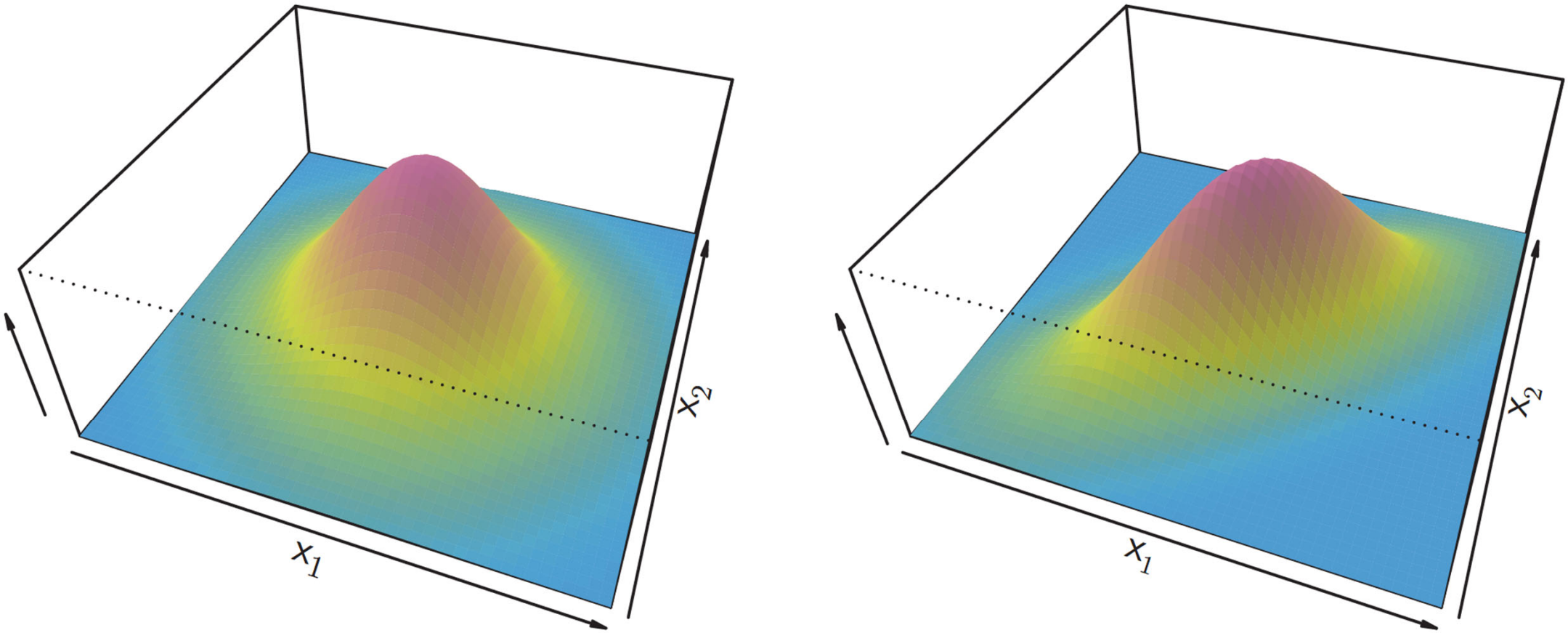
예측변수가 2개 이상일 때

- 다변량 정규분포(multivariate normal distribution) 가정
  - $X = (X_1, X_2, \dots, X_p)' \sim N(\mu, \Sigma)$. 즉,

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

  - μ: 평균벡터, Σ: 공분산행렬
- 가정: $X_k \sim N(\mu_k, \Sigma), k = 1,2, \dots, K$
- 베이즈 분류기
  - "$\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log \pi_k \ (k = 1,2, \dots, K)$"를 가장 크게 하는 범주로 분류 WHY?

그림4.5 이변량 정규분포. 왼쪽(무상관), 오른쪽(상관계수=0.7)



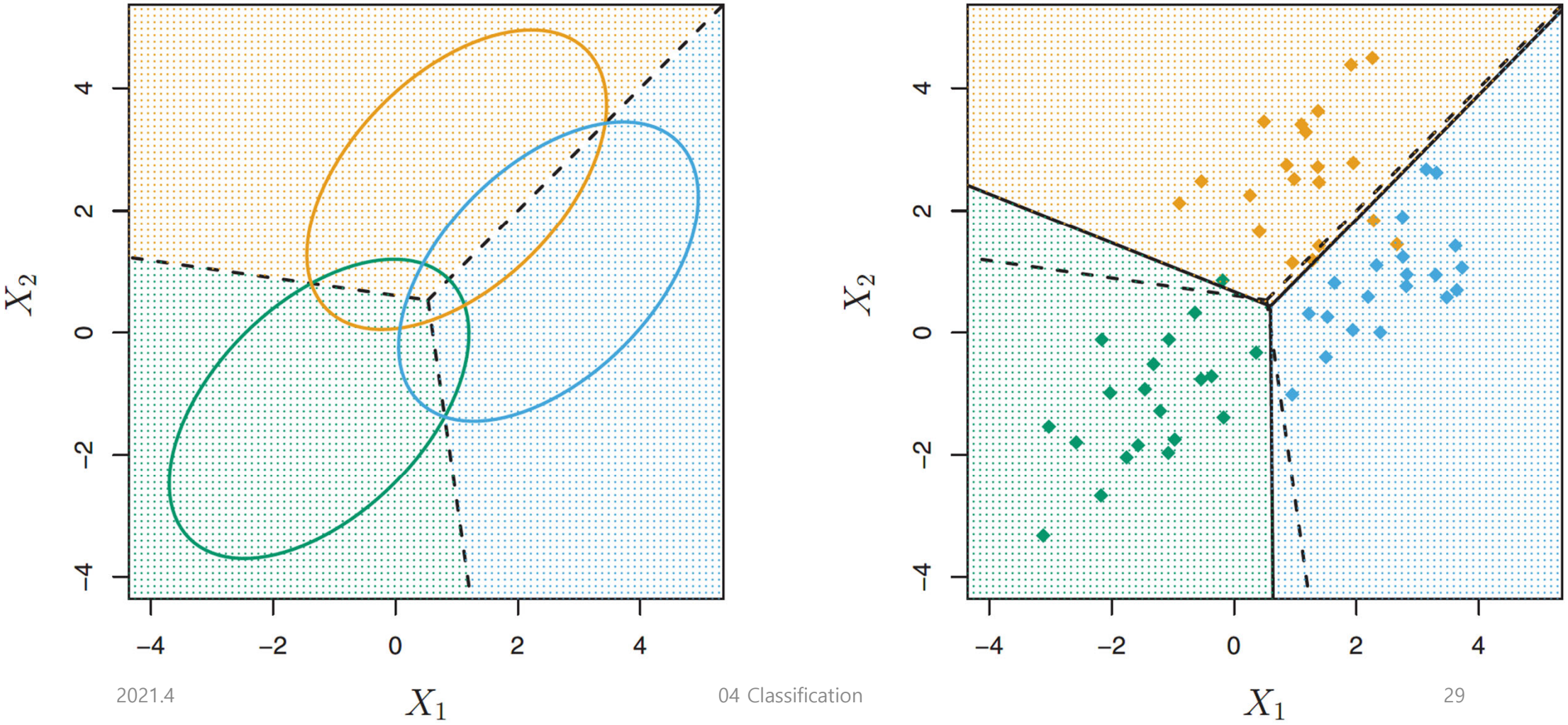
예측변수가 2개 이상일 때

- LDA 분류기
  - "$\hat{\delta}_k(x) = x'\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k\ (k = 1,2,\dots,K)$"를 가장 크게 하는 범주로 분류
  - $\hat{\pi}_k$: $\frac{n_k}{n}$, $\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$, $\hat{\Sigma}$: 복잡한 형태

가상 자료: $K = 3, p = 2$일 때

- 베이즈 오류율: 7.46%
- LDA 오류율: 7.7% (각 범주에서 $n = 20$)

그림4.6 평균 벡터는 다르고 공분산 행렬은 동일한 이변량 정규분포. 타원 안에 95% 자료가 포함됨. 20개의 랜덤표본 표시. 점선(베이즈 결정 경계), 실선(LDA 결정 경계)



채무불이행 자료:
학생 여부와 신용카드 빚

- 혼란(confusion)행렬: 반응변수의 실제 범주와 예측 범주의 개체수를 나타낸 표
- 훈련 오류율(혹은 정확도):
  - $\frac{252+53}{10000} = 0.0275$ $\left(\frac{9644+81}{10000} = 0.9725\right)$

두 가지 경고(caveats): 왜 오류율이 낮은가?

- n에 대한 p의 비율이 클수록 과대적합 (즉 훈련 오류율이 지나치게 낮음)에 노출될 수 있지만
- 채무 불이행 자료에서 반응변수 default의 "Yes" 비율이 3.33%로 매우 낮음. 따라서 null 분류기(즉 모두 "No"로 분류)와 큰 차이 없음

위양률 vs. 위음률

- 위양률(flase positive rate): 실제는 음성(“No”)인데 양성(“Yes”)으로 잘못 분류된 비율
  - $\frac{23}{9667} = 0.00238$ (매우 낮음)
- 위음률(false negative rate): 실제는 양성(“Yes”)인데 음성(“No”)으로 잘못 분류된 비율
- 신용카드사에서는 위음률이 작기 원함. 하지만 $\frac{252}{333} = 0.757$ (매우 높음)

임계값=0.5 일 때 혼란 행렬

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for* 23 *individuals who did not default and for* 252 *individuals who did default.*

민감도(sensitivity)와 특이도(specificity)

- 민감도: 양성("Yes")인 개체를 양성으로 바르게 분류하는 비율.
  "1 − 위음률"과 동일
  - $\frac{81}{333} = 0.243$ (낮음)
- 특이도: 음성("No")인 개체를 음성으로 바르게 분류하는 비율.
  "1 − 위양률"과 동일
  - $\frac{9644}{9667} = 0.998$ (매우 높음)

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

용어: 귀무가설(null), 대립가설(non-null)

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

용어: 재현율(recall), 정밀도(precision), 검정력(power)

민감도와 특이도

- 베이즈 분류기는 사후확률 0.5를 기준으로 나눔. 즉 "$P(\text{Yes}|X = x) > 0.5$" 이면 "Yes"로 분류! 통합 오류율을 가장 작게 하는데 목표를 둠
- 따라서 임계값(threshold)을 낮추면 위음률은 낮아짐(즉, 민감도는 증가함)

민감도와 특이도

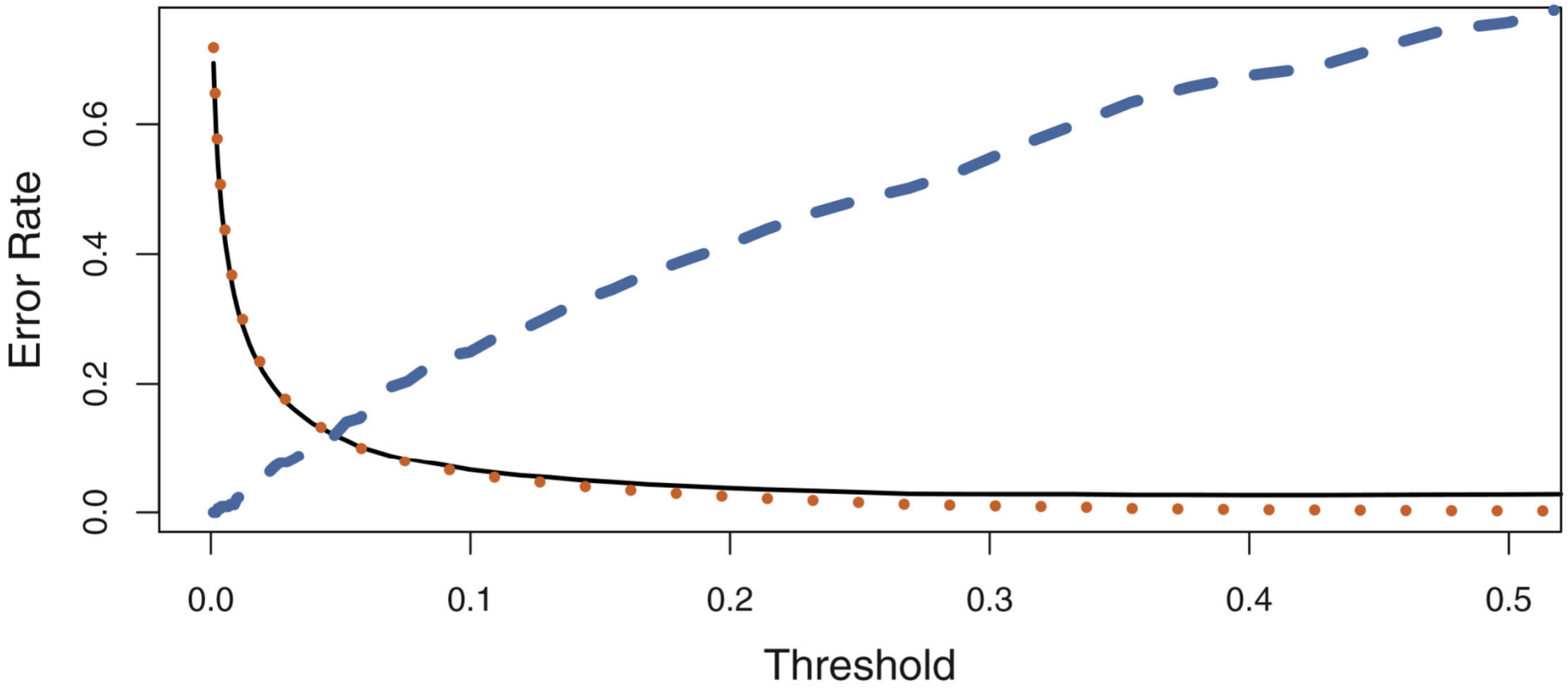
- 예: 채무불이행 자료에서 임계값이 0.2일 때,
  - 위음률: $\frac{138}{333} = 0.414$로 줄어듬. 그러나 위양률: $\frac{235}{9667} = 0.0243$, 통합 오류율: $\frac{138+235}{10000} = 0.0373$으로 늘어남
- 임계값은 얼마로 정해야 좋은가? 도메인 지식 이용!

임계값=0.2 일 때 혼란 행렬

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

TABLE 4.5. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds* 20 %.

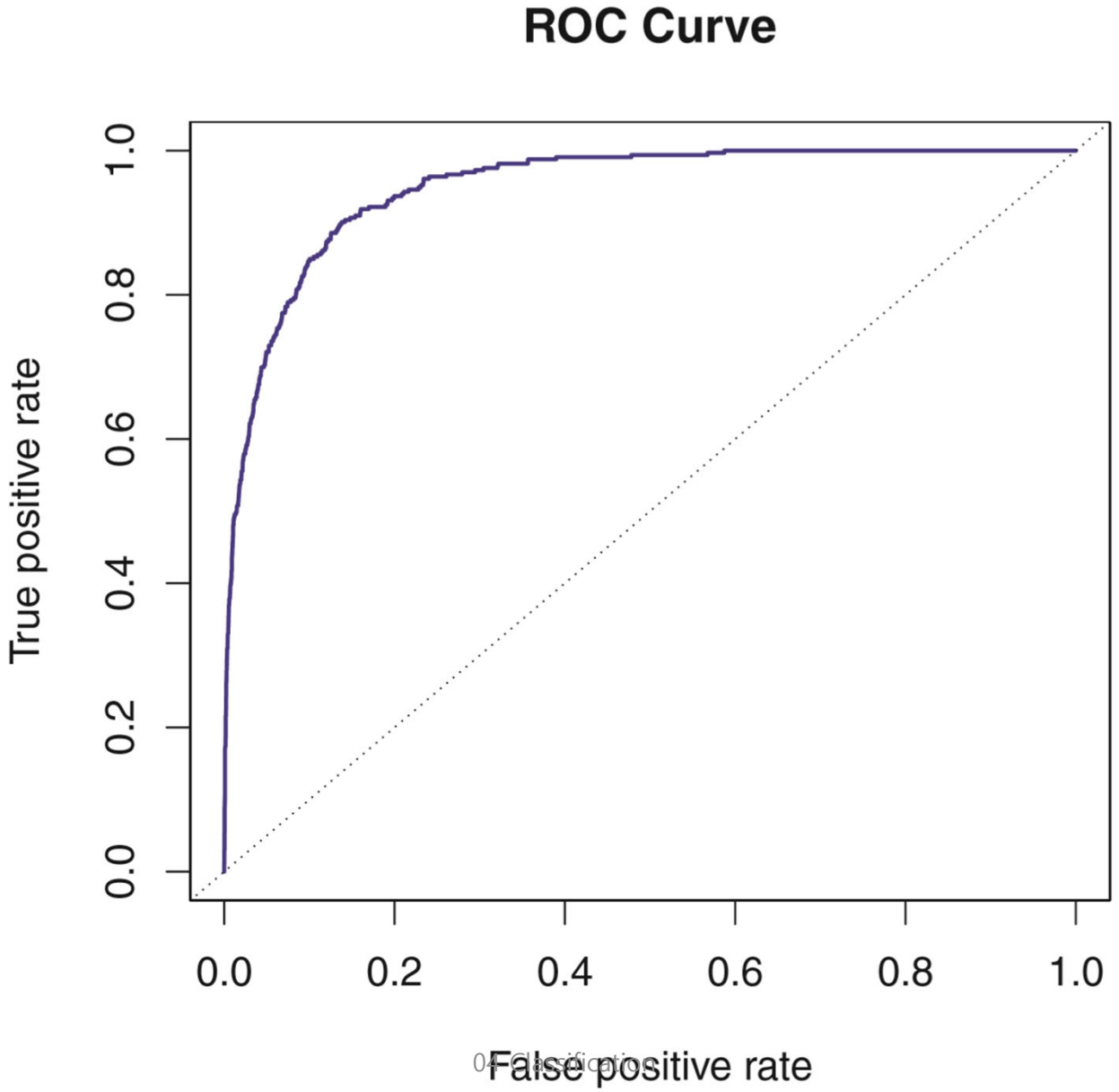
그림4.7 채무 불이행 자료. 파란색(위음률), 오렌지색(위양률), 검은색(통합 오류율)



ROC(receiver operating curve)와 AUC(area under curve)

- ROC: 임계값의 변화에 따라 위양률(즉 1 – 특이도)과 민감도를 동시에 나타낸 그림
  - 분류기들을 비교할 때 유용
- AUC: ROC 곡선 아래 넓이. "1"에 가까울수록 이상적인 방법. 0.5보다는 큼 WHY?

그림4.8 채무 불이행 자료. 점선(“no information” 분류기 사용. 즉, 학생 유무와 신용카드 빚이 채무 불이행 유무와 관련이 없을 때로 간주)



이차판별분석(quadratic discriminant analysis; QDA)

- 가정: $X_k \sim N(\mu_k, \Sigma_k), k = 1,2, \dots, K$
- 베이즈 분류기: $X = x$를 가진 개체에 대해 $\delta_k(x)$를 가장 크게 하는 범주로 분류

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$$

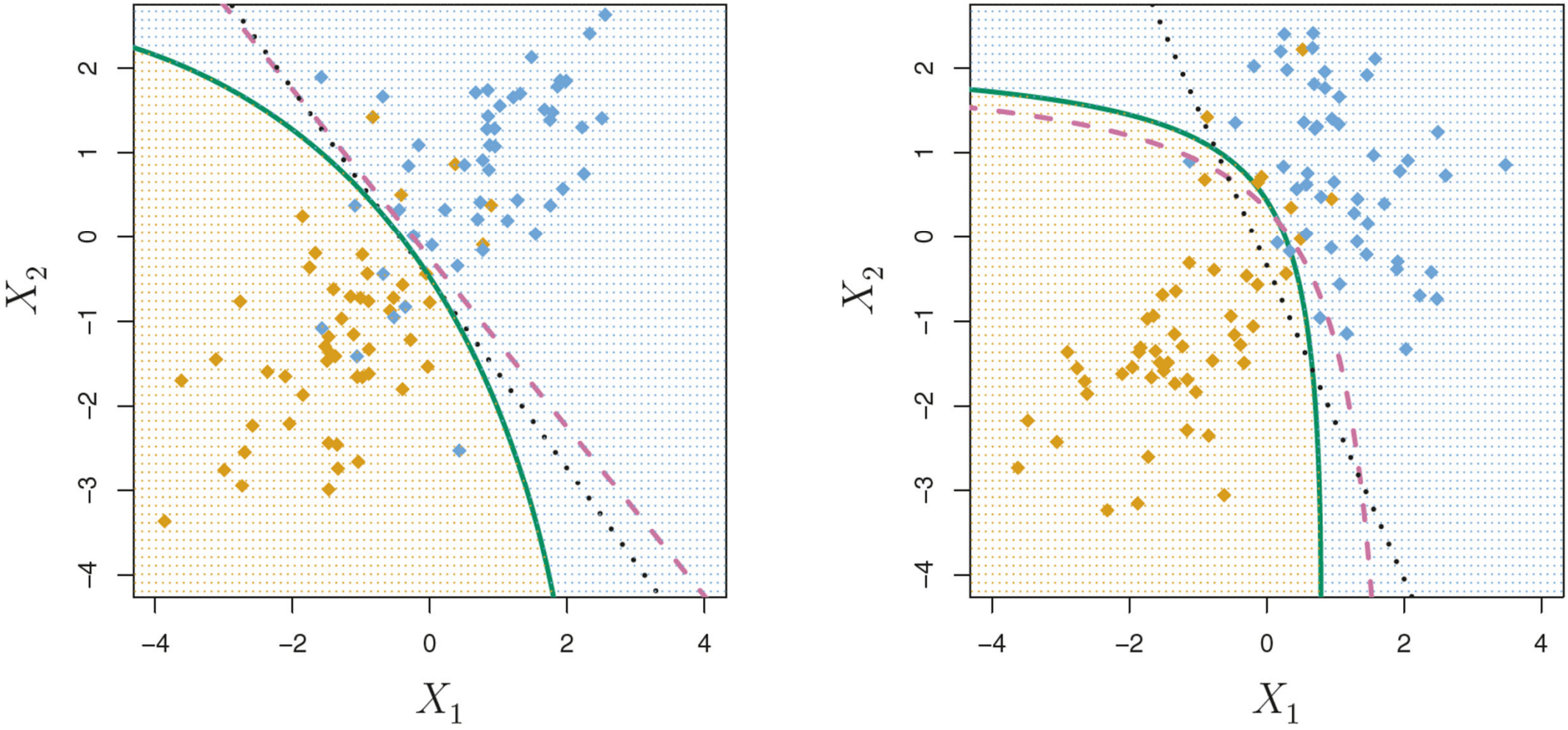
$$= -\frac{1}{2}x'\Sigma_k^{-1}x + x'\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$$

- QDA 분류기: 베이즈 분류기에서 $\mu_k \leftarrow \hat{\mu}_k$, $\Sigma_k \leftarrow \hat{\Sigma}_k$, $\pi_k \leftarrow \hat{\pi}_k$

LDA 분류기 vs. QDA 분류기

- 공분산행렬 추정: $\frac{p(p+1)}{2}$ vs. $K \times \frac{p(p+1)}{2}$
- QDA가 LDA보다 더 유연 ⇒ LDA의 분산은 더 작아지지만 편의는 더 커짐
- 자료수가 적으면 LDA를 선호하고, 자료가 많거나 범주간 공통 공분산행렬 가정이 맞지 않으면 QDA 선호

그림4.9 가상 자료. 왼쪽(공분산 행렬이 같은 경우), 오른쪽(공분산 행렬이 다른 경우). 보라색(베이즈 결정 경계), 검은색(LDA 결정 경계), 녹색(QDA 결정 경계). 왼쪽: LDA가, 오른쪽: QDA가 베이즈 결정 경계에 가까움



로지스틱 회귀모형 vs. LDA: $K = 2, p = 1$일 때

- $\log \frac{p_1(x)}{1-p_1(x)} = \beta_0 + \beta_1 x$ vs.
  $\log \frac{p_1(x)}{1-p_1(x)} = \log \frac{p_1(x)}{p_2(x)} = c_0 + c_1 x$ (c_0, c_1: μ_1, μ_2, σ^2의 함수) WHY?
- ML 추정량 이용 vs. 정규분포 가정. 표본평균과 표본분산 이용. 즉 추정방법만 다름. 결과는 유사
- 정규분포 가정이 맞으면 LDA가 우수하지만 그렇지 않으면 로지스틱 회귀모형이 우수
- 결정 경계는 선형(linear)

분류 방법들의 비교

- KNN: 비모수적 방법. 결정 경계가 비선형(non-linear)이지만, 예측만 가능하기 때문에 어떤 예측변수가 중요한지는 모름
- QDA: LDA(혹은 로지스틱 회귀모형)와 KNN 사이에 위치. 결정 경계가 이차 형태

분류 방법들의 비교

- 베이즈 결정 경계:
  - 시나리오 1, 2, 3: 선형
  - 시나리오 4, 5, 6: 비선형(이차 혹은 더 복잡한 형태)
- 100번 반복

분류 방법들의 비교

- 가정: $K = 2, p = 2, n_k = 20\ (k = 1,2)$
- 시나리오 1: X_k ~bivariate normal. $\rho = 0$
- 시나리오 2: X_k ~bivariate normal. $\rho = -0.5$
- 시나리오 3: X_k ~bivariate $t. \rho = 0, n_k = 50$
  - LDA의 정규성 가정 위배
- 결론: 결정 경계가 선형이면 LDA 혹은 로지스틱 회귀모형이 우수

그림4.10 선형 시나리오 경우 테스트 세트 오류율. KNN-1은 K=1, KNN-CV은 교차검증(cross-validation)으로 결정된 K의 값을 적용

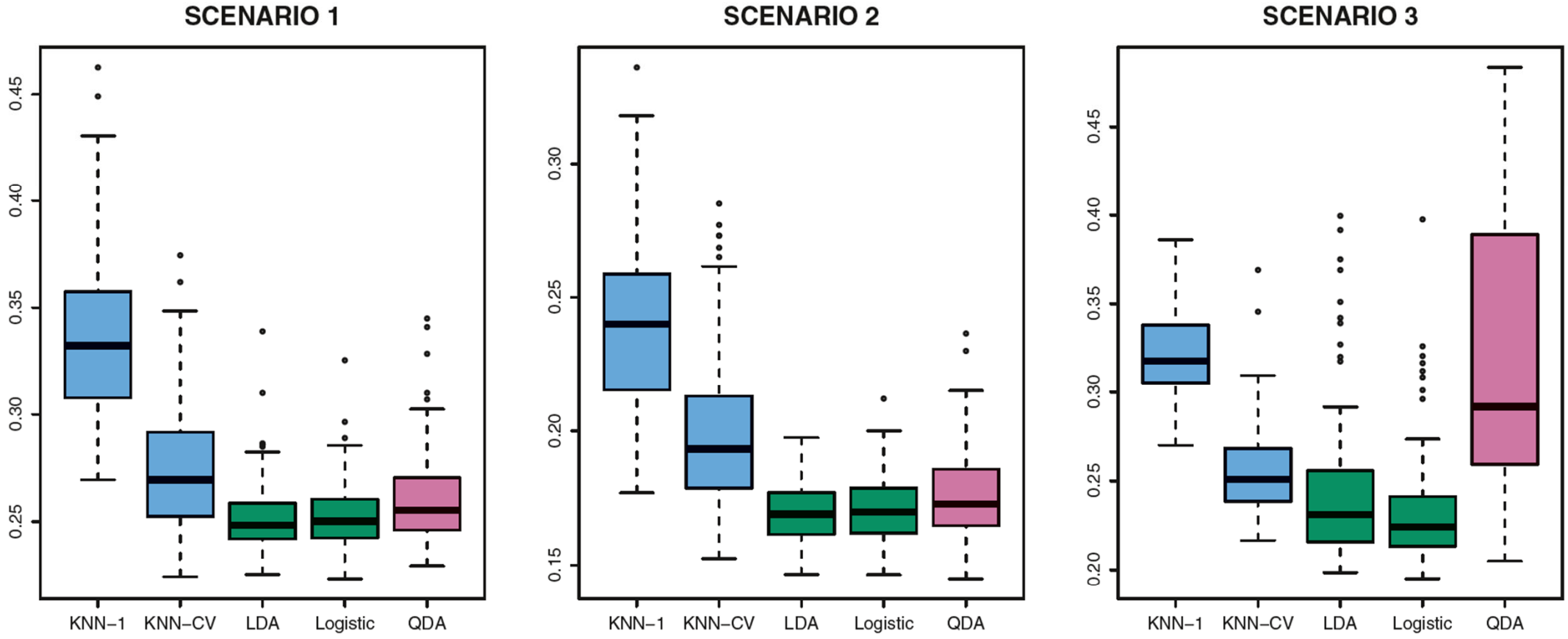


FIGURE 4.10. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

분류 방법들의 비교

- 시나리오 4: X_k ~bivariate normal. $\rho_1 = 0.5, \rho_2 = -0.5$
- 시나리오 5: X_k ~bivariate normal. $\rho = 0$, Y ~로지스틱 함수 (이때 예측변수는 $X_1^2, X_2^2, X_1 \times X_2$)
  - 결정 경계는 이차 형태
- 시나리오 6: X_k ~bivariate normal. $\rho = 0, Y$ ~비선형 함수
  - 결정 경계는 (이차보다) 복잡한 형태
- 결론: 결정 경계가 대체로 비선형이면 QDA가 우수. 매우 복잡하면 KNN이 우수. 다만 평활(smoothness) 수준 선택 필요

그림4.11 비선형 시나리오 경우 테스트 세트 오류율. KNN-1은 K=1, KNN-CV은 교차검증(cross-validation)으로 결정된 K의 값을 적용

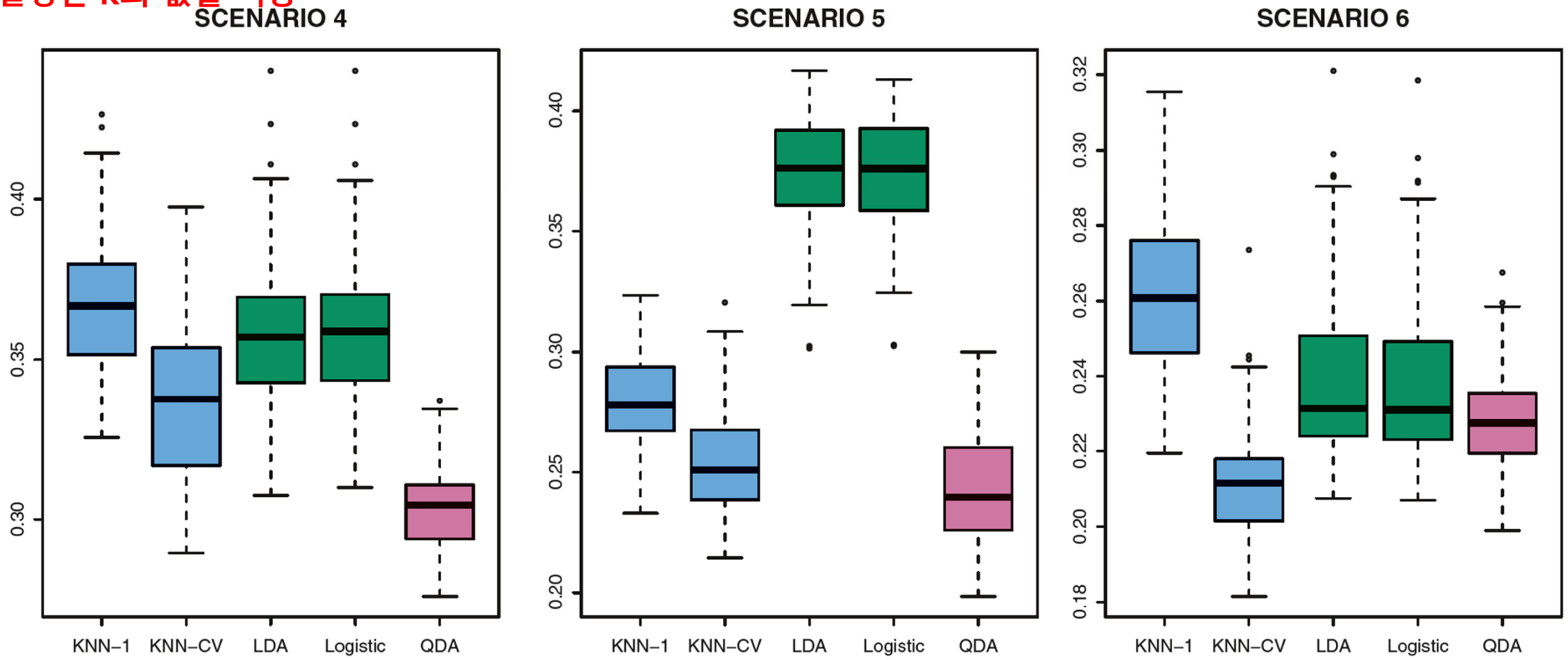


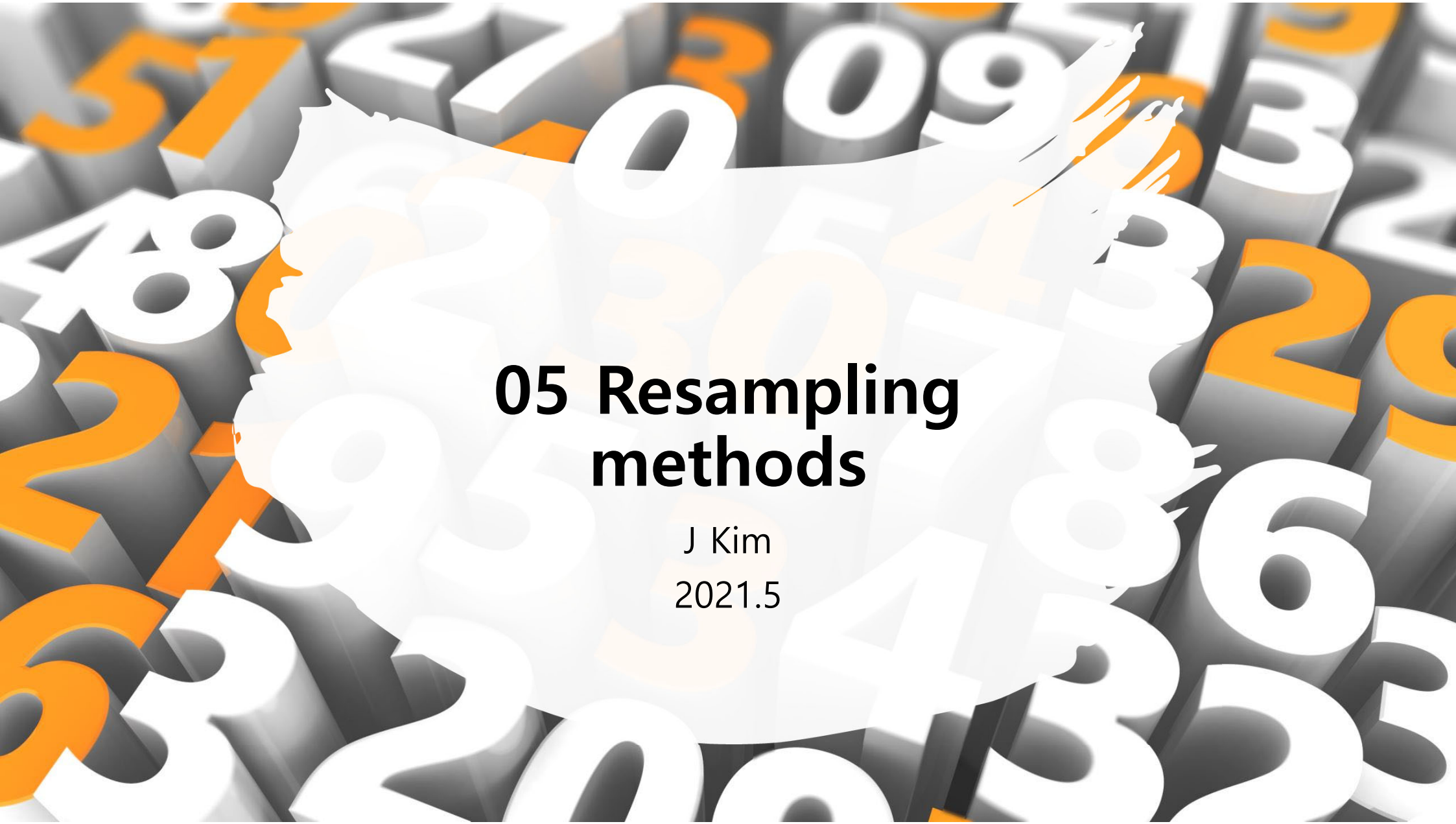
FIGURE 4.11. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*

과제(5월25일마감)

- 연습문제 4장: 3, 7, 9, 11

Thank you!

Move on to 05 Resampling methods



05 Resampling methods

J Kim

2021.5

교차검증
부스트랩

Outline

재표집(resampling)

- 재표집이란? 훈련세트에서 반복적으로 샘플을 뽑고 그 샘플에 관심 있는 모형을 적합하는 것
- 계산량이 많지만 컴퓨팅 능력과 속도가 향상되어 가능해짐
- 대표적으로 교차검증(cross-validation, CV)과 부스트랩(bootstrap) 방법이 있음
- CV는 모형 평가(model assessment)와 모형 선택(model selection)을 위해 테스트 오류율을 추정
- 부스트랩은 추정량의 정확도(accuracy)를 측정하는데 사용

검증세트(validation set) 방법

- 랜덤하게 관찰 개체를 훈련(training)세트와 검증(validation)세트(혹은 hold-out 세트)로 나눔
- 훈련세트로 모형선택 후 검증세트로 MSE 계산
- 검증세트 MSE $= \frac{1}{n_v}\sum_{i\in V_s}(y_i - \hat{y}_i)^2$
  - $n_v = \#(V_s), V_s$: 검증세트에 속하는 개체들의 인덱스

검증세트 방법: 약점

- 검증세트가 랜덤하게 결정되므로 검증세트에 따라 MSE가 변할 수 있음
- 자료 중 일부만으로 추정하기 때문에 검증세트 오류율이 자료 전체를 써서 추정한 모형의 테스트 오류율보다 클(overestimate) 수 있음

그림5.1 전체 개체를 랜덤하게 둘로 나눔. 파란색(훈련 세트), 오렌지색(검증 세트)

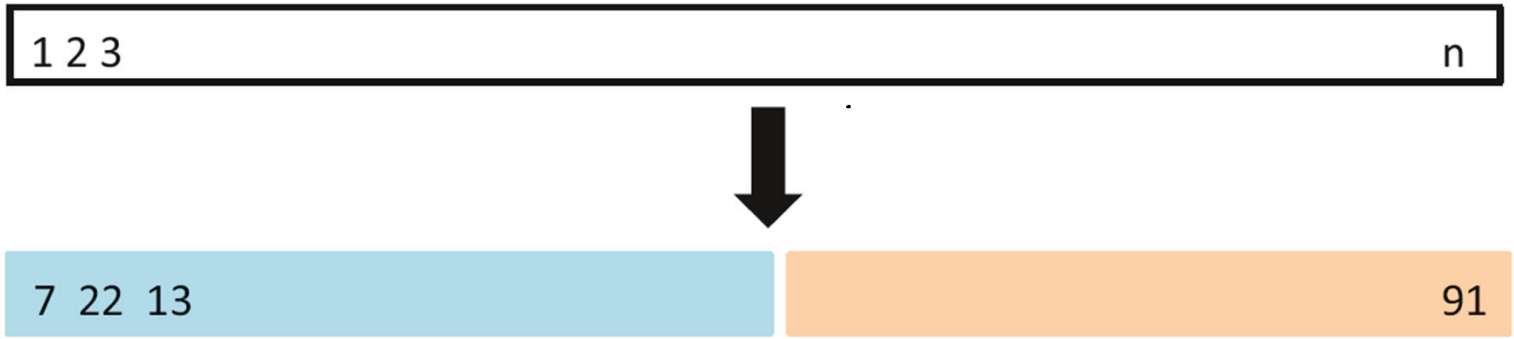


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

그림5.2 자동차 자료. 마력만 예측변수로 잡고 다항회귀모형을 적합. 왼쪽(검증 세트 MSE), 오른쪽(10번 반복. 검증 세트 MSE 계산. 변동이 큼)

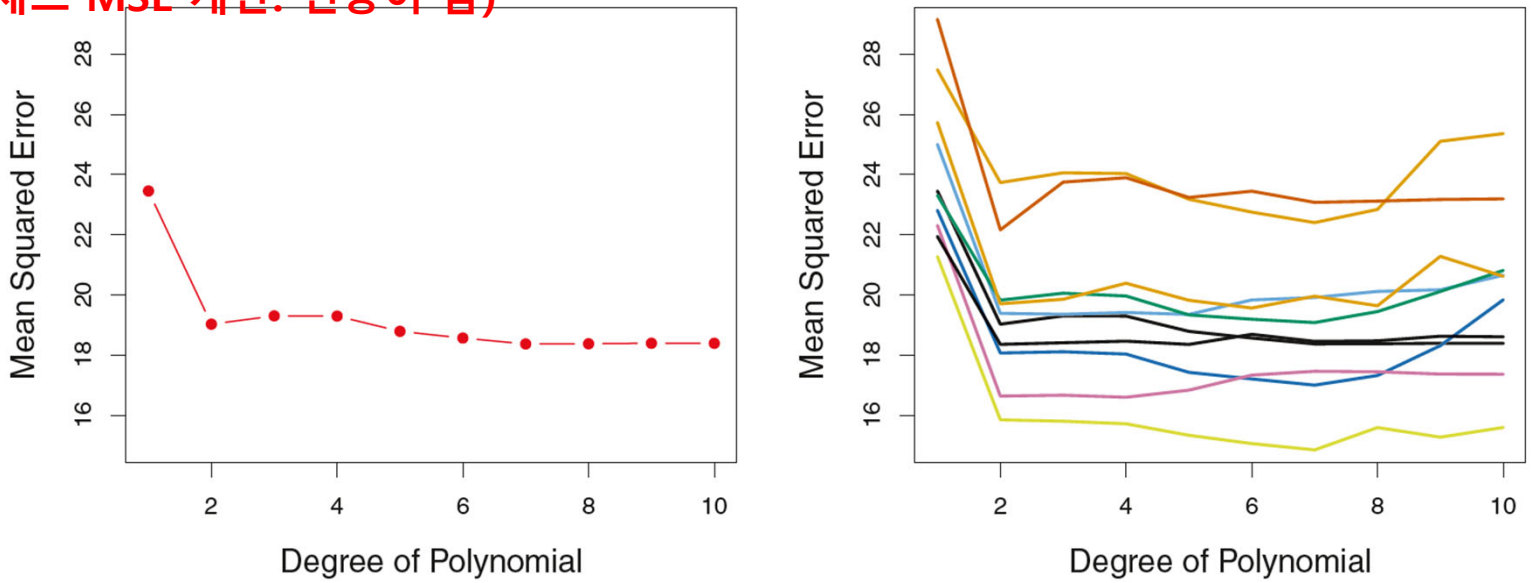


FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

일대다 교차검증(leave-one-out CV, LOOCV)

- 검증세트방법처럼 관찰 개체를 2개 그룹으로 나눔. 한 그룹은 1개 개체로만 만들고 다른 그룹은 나머지 $(n-1)$개 개체로 만듦
- 첫 번째 개체 (x_1, y_1)를 검증세트, 나머지 $\{(x_2, y_2), \dots, (x_n y_n)\}$를 훈련세트로 놓고 모형을 선택 후,

  첫 번째 개체의 MSE를 계산 $\Rightarrow \mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$
- 같은 방법으로 $\mathrm{MSE}_2, \dots, \mathrm{MSE}_n$
- LOOCV MSE: $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$

일대다 교차검증

- Remarks 1: 검증세트 방법보다 훈련세트 크기가 커 편의가 줄고 테스트 오류율을 과대추정 하지 않음
- Remarks 2: 검증세트 방법처럼 훈련/검증세트가 랜덤 하지 않아서 MSE가 항상 동일! (그림5-4 왼쪽 참조)
- Remarks 3: LR에서는 n번 반복하지 않고서도 $\mathrm{CV}_{(n)}$ 계산 가능.

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

그림5.3 파란색(훈련 세트), 오렌지색(테스트 세트). 1개 개체로 된 n개의 폴드를 만듦.

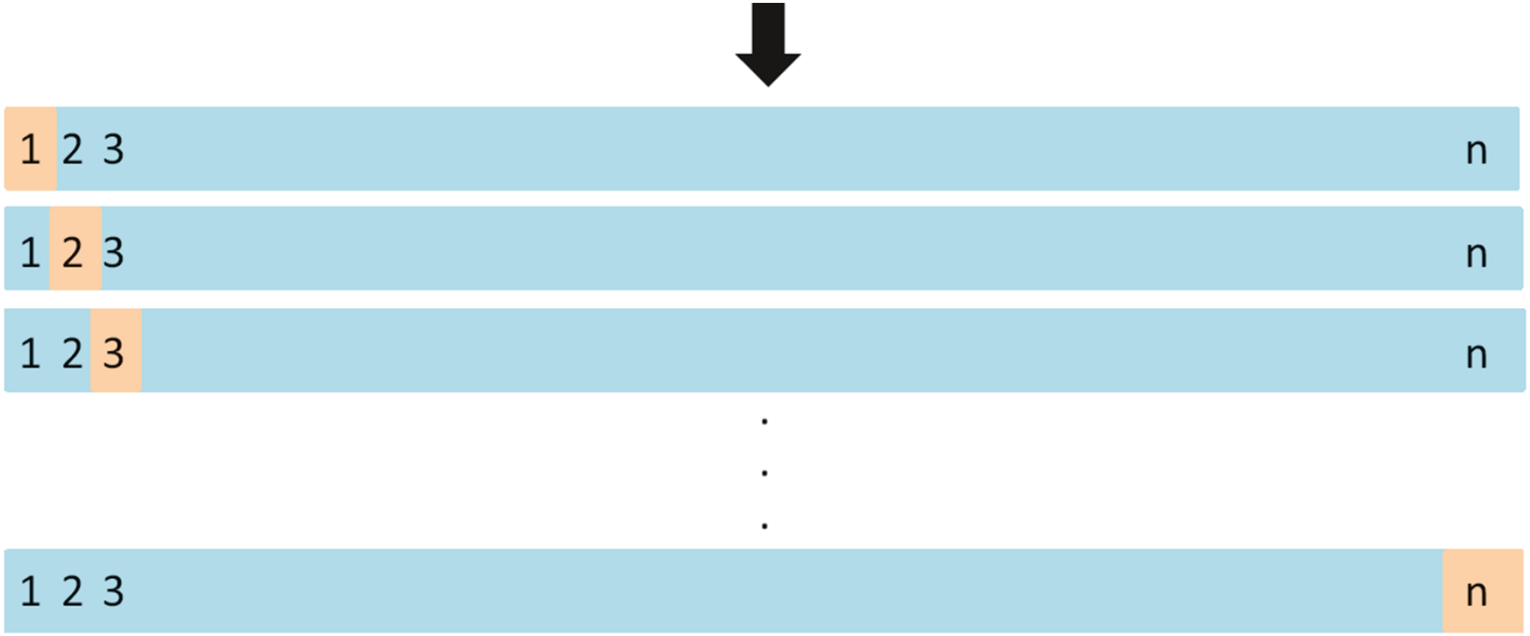


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation*

그림5.4 자동차 자료. 마력만 예측변수로 잡고 다항회귀모형을 적합. 왼쪽(LOOCV MSE), 오른쪽(10-폴드 CV 방법을 9번 반복. 변동이 작음)

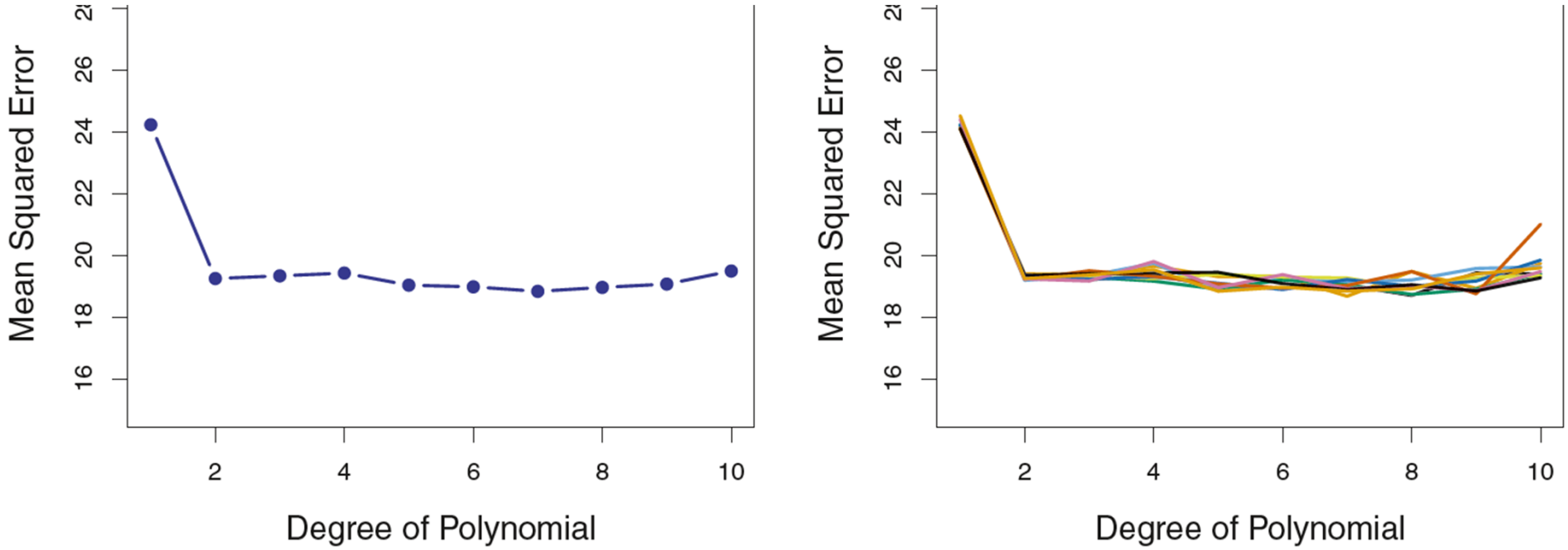


FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10-*fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

k-폴드 교차검증(k-fold CV)

- 전체 개체를 k개 폴드(fold)로 나눔. 즉 $C_1, C_2, \dots, C_k$
- $C_j (j = 1,2, \dots, k)$: $j-$번째 폴드에 포함된 개체들의 인덱스
- 첫 번째 폴드(hold-out 폴드)는 검증 세트, 나머지 $(k-1)$개 폴드는 훈련 세트로 놓고 모형을 선택한 후 첫 번째 폴드의 MSE 계산
  - $\mathrm{MSE}_1 = \frac{1}{n_1}\sum_{i \in C_1}(y_i - \hat{y}_i)^2$, $n_1 = \#(C_1)$
- 같은 방법으로 $\mathrm{MSE}_2, \dots, \mathrm{MSE}_k$
- k-폴드 CV MSE: $\mathrm{CV}_{(k)} = \sum_{j=1}^{k} \frac{n_j}{n} \mathrm{MSE}_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i \in C_j}(y_i - \hat{y}_i)^2$

그림5.5 전체 개체를 5개의 폴드로 나눔. 파란색(훈련 세트), 오렌지색(테스트 세트).

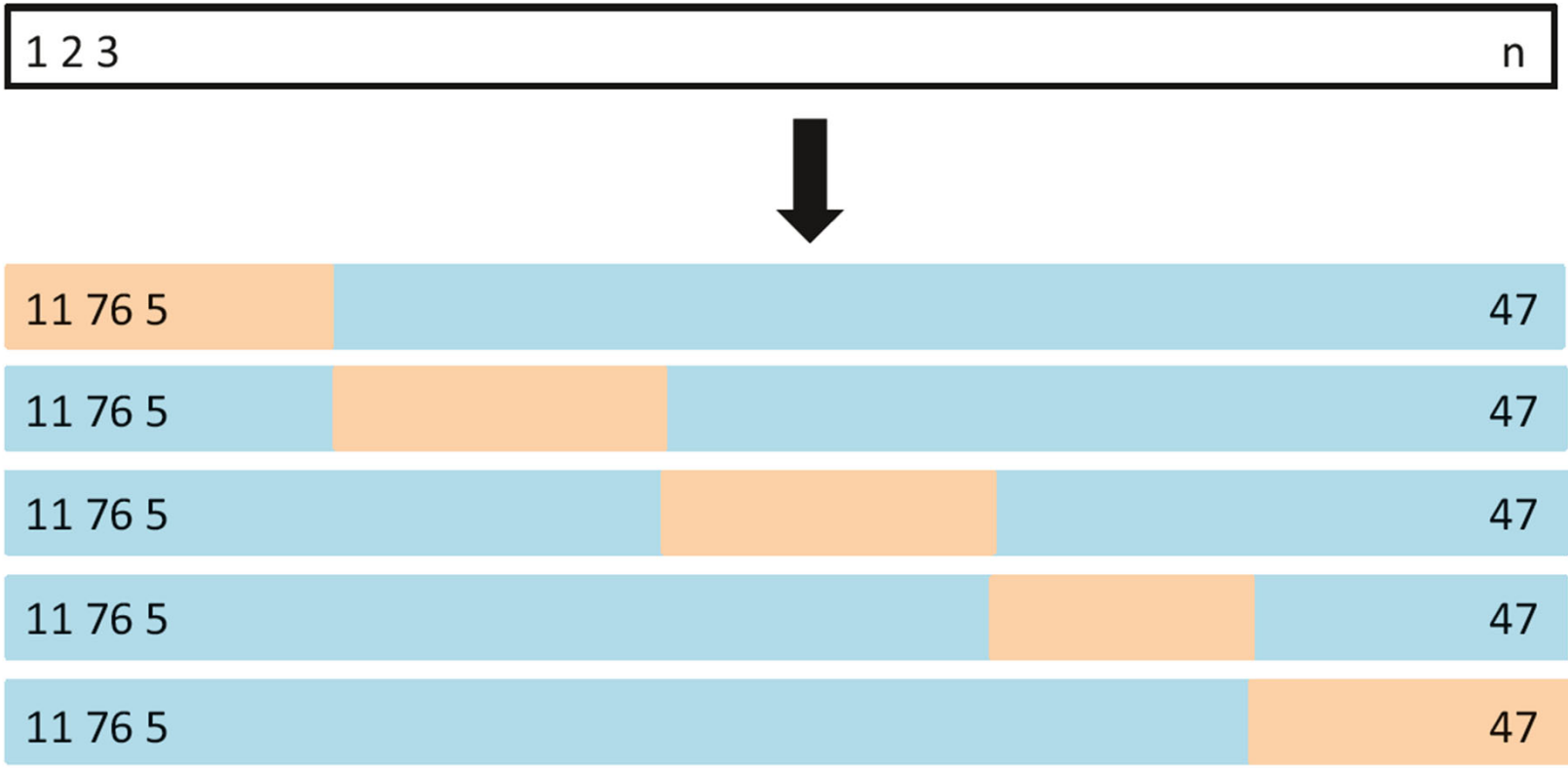


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

k-폴드 교차검증

- Remarks 1: LOOCV는 k-폴드 CV의 특수한 경우. 즉 $k = n$인 경우에 해당
- Remarks 2: k -폴드 CV의 계산양이 LOOCV보다 적음. 다만 LR은 예외
- Remarks 3: k -폴드 CV에 의한 MSE는 CV에 따라 변동하지만 (그림5-4 오른쪽 참조) 검증세트방법보다 변동폭이 작음 (그림5-2 오른쪽 참조)

k-폴드 교차검증

- Remarks 4: 가상 자료(그림2-9, 2-10, 2-11)에 대한 LOOCV와 10-폴드 CV의 MSE와 테스트 MSE의 비교 (그림5-6 참조)
- Remarks 5: True 테스트 MSE를 가장 작게 하는 유연성 ≈ CV MSE를 가장 작게 하는 유연성. 따라서 유연성 수준을 결정하는 데 CV MSE를 적용하는 것은 적절!

그림5.6 가상 자료(그림2.9-2.11). 파란색: True 테스트 MSE, 검은색: LOOCV MSE, 오렌지색: 10-CV MSE. 교차 표시는 MSE가 가장 작은 값

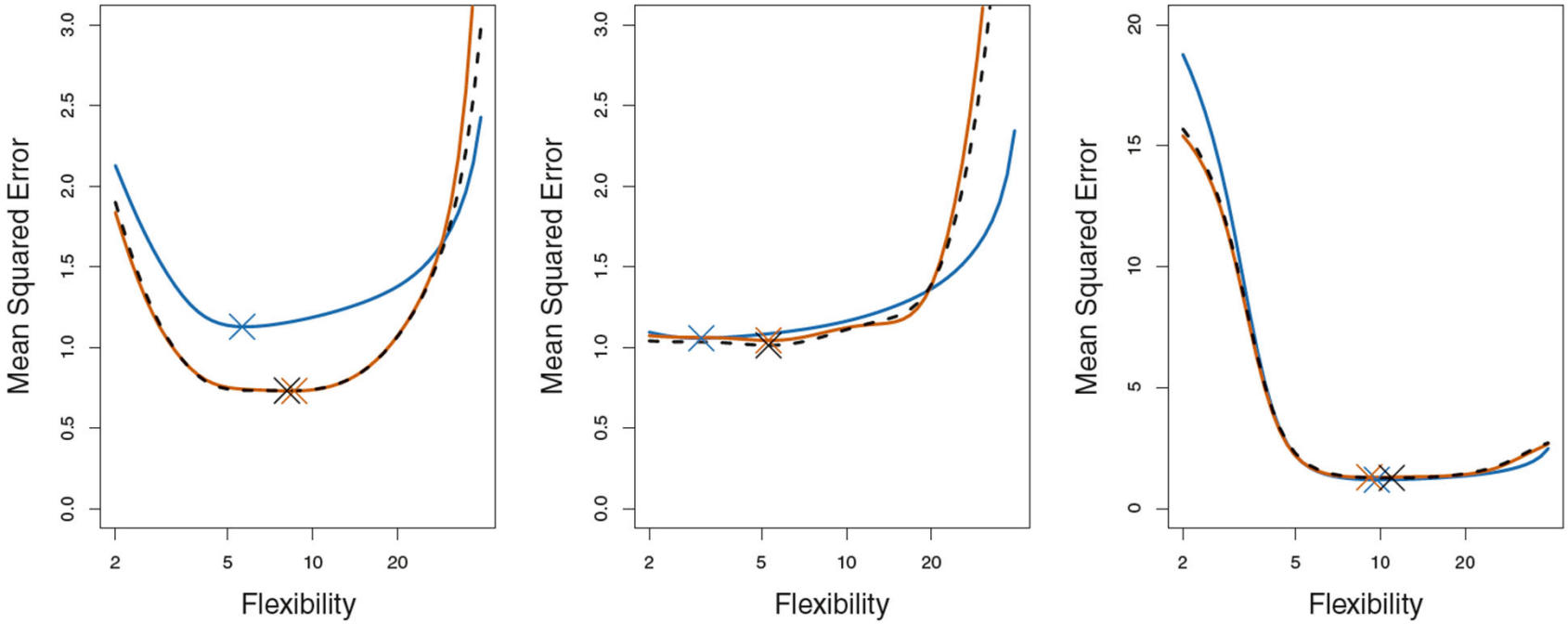


FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9 (*left*), 2.10 (*center*), and 2.11 (*right*). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the* 10*-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

k-폴드 CV에서 편의와 분산의 교환

- k-폴드 CV는 LOOCV보다 테스트 MSE를 더 정확하게 추정할 수 있을까?
  - 편의 측면에서는 LOOCV가 k-폴드 CV보다 유리
  - LOCOV에서 n개 훈련세트들은 k-폴드 CV에서 k개 훈련세트들보다 더 유사함
  - 일반적으로 강하게 상관된 측도들의 평균은 덜 상관된 측도들의 평균보다 분산이 큰 경향이 있음
  - LOOCV의 분산이 k-폴드 CV보다 큼
- Yes! But, k-폴드 CV에서 k의 결정은 편의도, 분산도 극단적으로 크지 않도록. 보통 $k = 5, 10$

분류문제

- MSE 대신에 오류율!
- 검증세트방법: 검증세트의 오류율 계산
- LOOCV: $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i \, , \mathrm{Err}_i = I(y_i \neq \hat{y}_i)$
- k -폴드 CV: $\mathrm{CV}_{(k)} = \sum_{j=1}^{k} \frac{n_j}{n} \mathrm{Err}_j = \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in C_j} I(y_i \neq \hat{y}_i)$
  - $\mathrm{Err}_j = \frac{1}{n_j}\sum_{i \in C_j} I(y_i \neq \hat{y}_i)$: $j-$번째 폴드의 오류율

분류문제

- Remarks 1: CV 오류율은 다소 과소추정하지만 테스트 오류율과 흡사 (그림5.8 참조)
- Remarks 2: 검증 세트(폴드)의 개수(k)를 고정했을 때,
  - 로지스틱 회귀모형에서 예측변수의 차수 혹은
  - k –최근접이웃 방법에서 k를 결정하는 데 CV 오류율을 사용하는 것은 적절!

그림5.7 가상 자료(그림2.13, 예측변수: X_1, X_2; K=2).

보라색: 베이즈 결정 경계.
검은색: 로지스틱 회귀모형을 적합.
왼쪽 상단: 선형 (X_1, X_2),
오른쪽 상단: 2차 다항회귀 $(X_1, X_2, X_1X_2, X_1^2, X_2^2)$,
왼쪽 하단: 3차 다항회귀
$(X_1, X_2, X_1X_2, X_1^2, X_2^2, X_1^2X_2, X_1X_2^2, X_1^3, X_2^3)$,
오른쪽 하단: 4차 다항회귀
$(X_1, X_2, X_1X_2, X_1^2, X_2^2, X_1^2X_2, X_1X_2^2,$
$X_1^3, X_2^3, X_1^3X_2, X_1X_2^3, X_1^2X_2^2, X_1^4, X_2^4)$.

베이즈 오류율=0.133.
테스트 오류율=0.201, 0.197, 0.160, 0.162

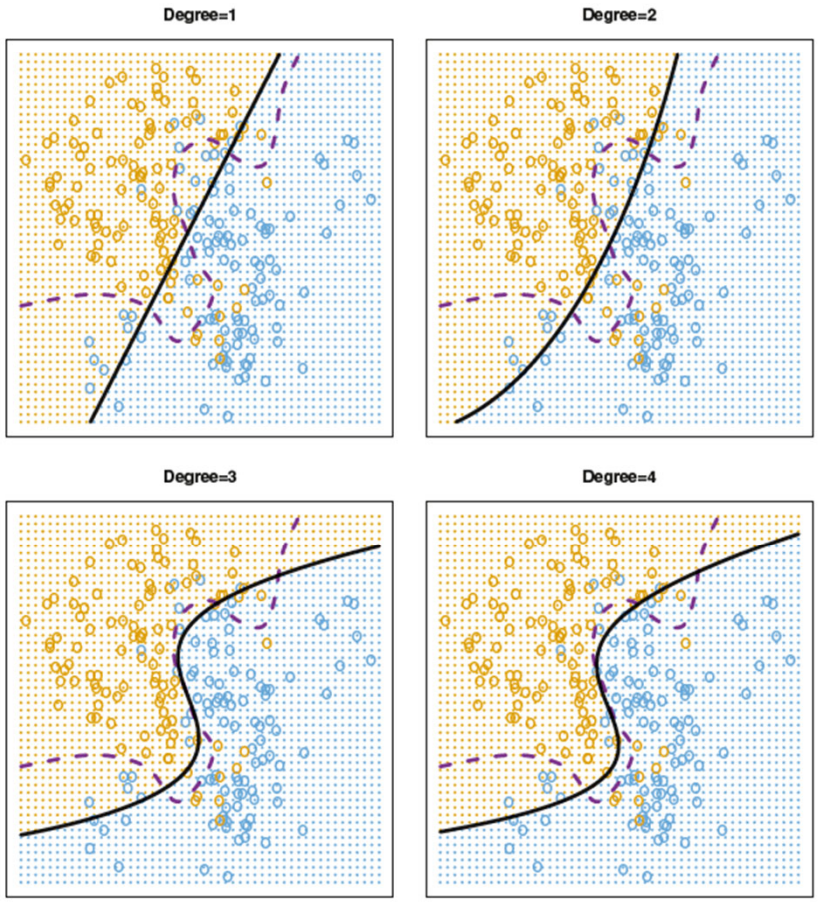


FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in 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.

그림5.8 가상 자료(그림2.13). 오렌지색: 테스트 오류율, 파란색: 훈련 오류율, 검은색 : 10-CV 오류율. 교차 표시는 오류율이 가장 작은 값. 왼쪽: 로지스틱 다항회귀, 오른쪽: KNN

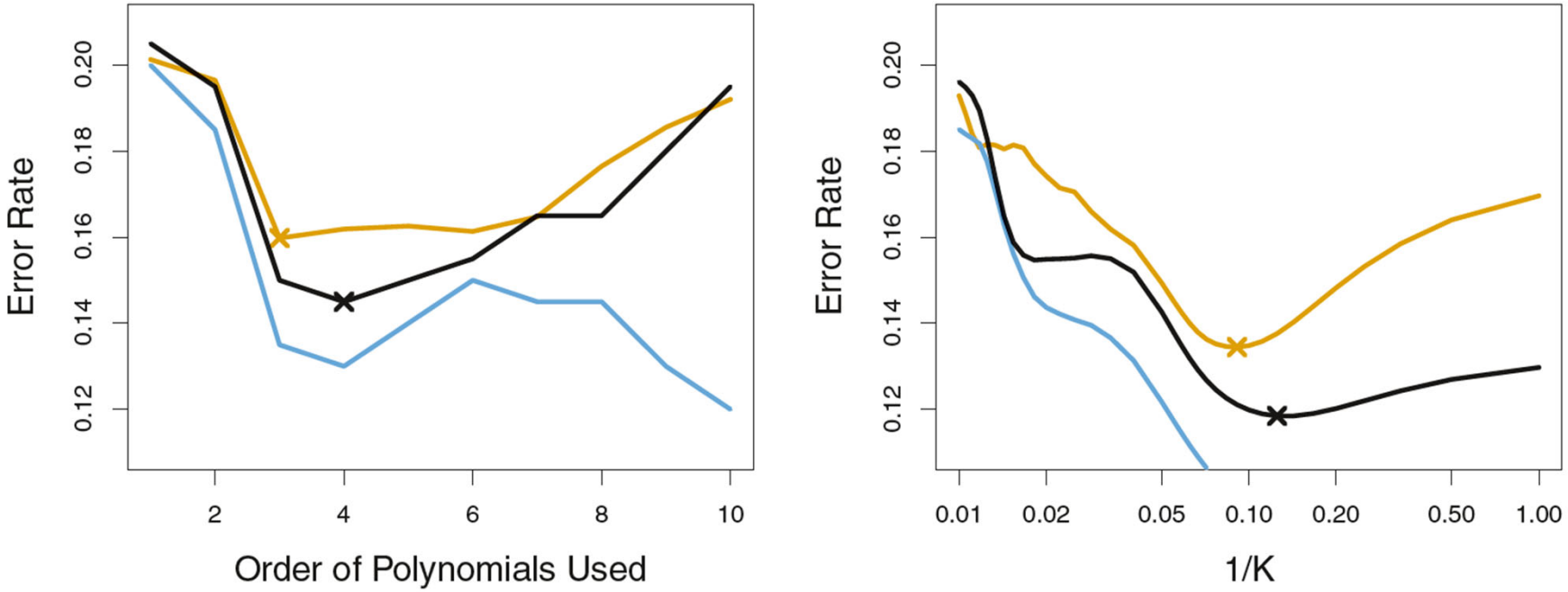


FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

부스트랩(bootstrap)

- 언제 유용한가? 추정량의 정밀도(accuracy)(혹은 불확실성, uncertainty)를 수량화 하고자 할 때

토이 예제

- 일정한 금액으로 두 곳에 투자하려 함
- 두 투자처의 이득을 각각 X, Y라 할 때, 어떻게 배분하면 총이득의 분산을 가장 작게 할 수 있을까? 즉

$$Var(\alpha X + (1-\alpha)Y)$$

을 가장 작게 하는 α의 값은 얼마인가?

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

- 주어진 자료로부터 α를 추정: $\hat{\alpha} = \frac{\widehat{\sigma_Y^2} - \hat{\sigma}_{XY}}{\widehat{\sigma_X^2} + \widehat{\sigma_Y^2} - 2\hat{\sigma}_{XY}}$

모의실험자료

- $\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5 \Rightarrow \alpha = 0.6$
- 100 쌍
- 1000번 반복: $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$
  - $\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996$
  - $\mathrm{SE}(\hat{\alpha}) \approx \sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$

그림5.9 X, Y 투자 자료(100개). 추정량 = 0.576, 0.532, 0.657, 0.651(좌에서 우, 위에서 아래순으로)

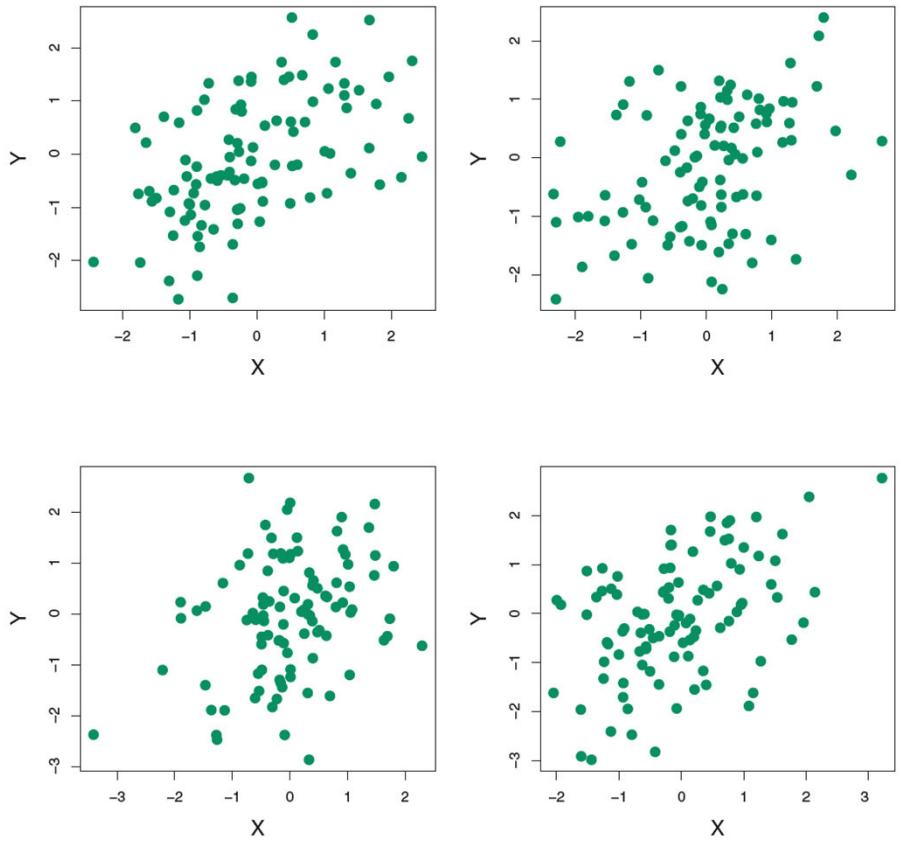


FIGURE 5.9. *Each panel displays* 100 *simulated returns for investments* X *and* Y. *From left to right and top to bottom, the resulting estimates for* α *are* 0.576, 0.532, 0.657, *and* 0.651.

그림5.10 왼쪽: 1000개의 가상 자료를 써서 구한 $\hat{\alpha}$의 히스토그램. 중앙: 1000개의 부스트랩 자료를 써서 구한 $\hat{\alpha}$의 히스토그램. 오른쪽: 가상 자료와 부스트랩 자료의 결과에 대한 상자그림. 분홍색은 α의 True 값 = 0.6

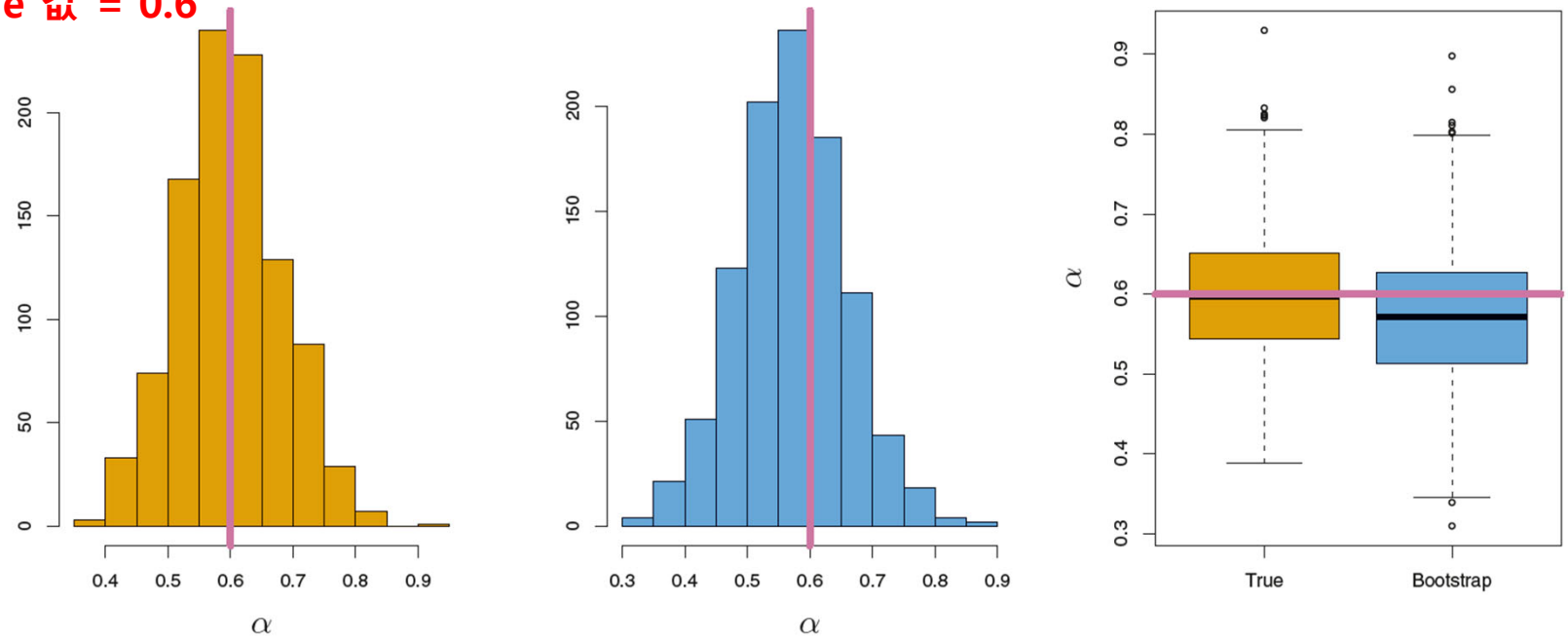


FIGURE 5.10. Left: *A histogram of the estimates of* α *obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of* α *obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of* α *displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of* α.

부스트랩

- 부스트랩: 새로운 자료를 모집단에서 다시 추출할 수 없기 때문에 실제 자료에서 개체들을 반복적으로 뽑아 새로운 자료를 생성
- 새로운 자료에는 같은 개체가 여러 번 중복될 수도 있지만 서로 독립인 개체로 취급

그림5.11 n=3 일 때 원 자료에서 부스트랩 표본을 얻고 $\hat{\alpha}$ 계산

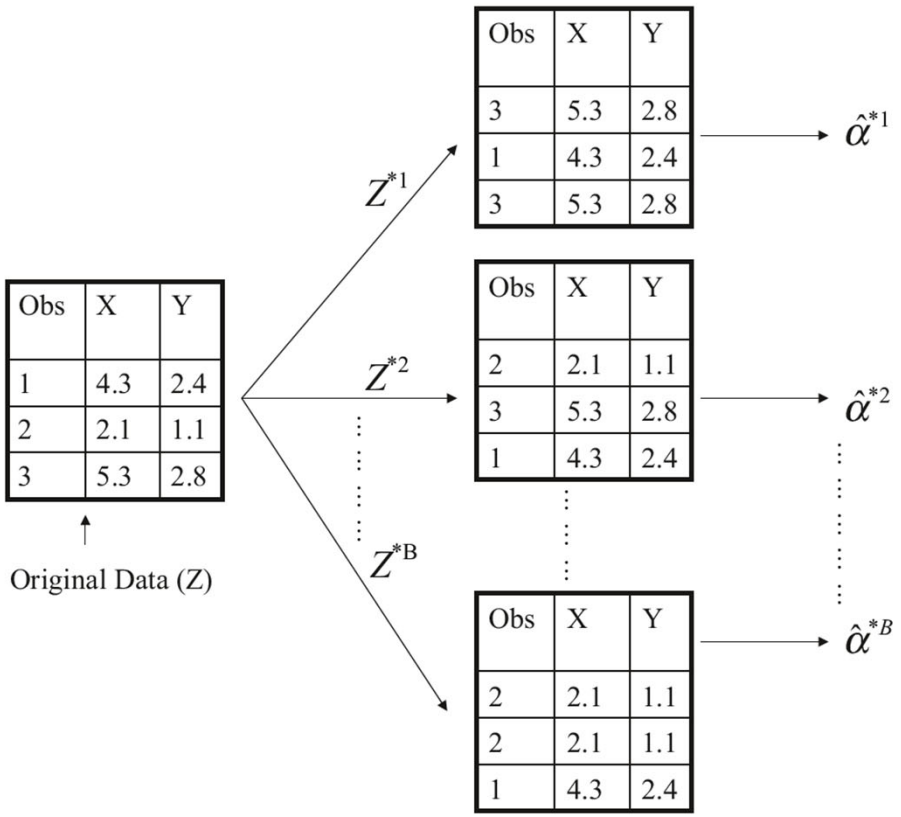


FIGURE 5.11. *A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.*

부스트랩

- $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$: B개 서로 다른 부스트랩 자료
- $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$: (부스트랩 자료를 써서 구한) α의 추정량
- 부스트랩 추정량의 표준오차:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \bar{\alpha}_B)^2}$$

  - $\bar{\alpha}_B = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}$
- 예: $B = 1000$, $SE_B(\hat{\alpha}) = 0.087$

과제(6월1일마감)

- 연습문제 5장: 2, 6, 7, 9

Thank you!

Move on to 06 Linear model section and regularization: Part I

06 Linear model selection & regularization: Part I

J Kim

2020.6

Outline

부분집합 선택: 최적 부분집합 선택, 단계적 선택
수축: 릿지회귀, 라쏘
차원 축소: 주성분회귀
고차원자료

선형모형에서 최소제곱방법을 개선할 다른 방법은 없는가?

- 왜 개선하려고 하는가?
  - 예측변수와 반응변수 간에 선형관계가 있다면 LS 방법은 편의가 작고, $n \gg p$라면 LS 방법은 분산이 작고 테스트 MSE도 작게 됨 ⇒ 그러나 n이 p보다 충분히 크지 않다면 LS 방법은 변동성이 커져 과대적합(overfitting)에 빠질 수도! 특히 $n < p$라면 LS 방법은 유일해가 존재하지 않고 분산이 무한대가 되어 적용 불가능해짐
  - LM에 포함된 예측변수 중에는 반응변수와 연관되지 않은(irrelevant) 것도 있음 ⇒ 모형을 복잡하게 함

선형모형에서 최소제곱방법을 개선할 다른 방법은 없는가?

- 답: 예측력(prediction accuracy)과 해석력(model interpretability)을 모두 향상시킬 수 있음
- 세 가지 방법
  - 부분집합 선택(subset selection): 반응변수와 연관된 예측변수의 부분집합을 찾아낸 후 LS 방법을 적용
  - 수축(shrinkage) 혹은 규제(regularization): 모든 예측변수를 모형에 포함시키지만 LS 방법에 비해 회귀계수 추정값이 0 쪽으로 쪼그라들게 하여 분산을 줄이고 변수선택(variable selection)을 수행
  - 차원 축소(dimension reduction): p-차원 예측변수를 $M(<p)$-차원으로 낮추고 선형결합에 LS 방법을 적용

최적 부분집합 선택(best subset selection)

- 예측변수들의 모든 조합($2^p, p$: 예측변수 개수)에 대한 회귀모형을 적합시킴
- 두 스텝으로 나뉨
  - 같은 크기의 부분집합 별로 최적(best) 모형을 선택
  - $(p+1)$개의 모형 중에서 최적 모형을 선택

최적 부분집합 선택: 알고리즘

- Step 1: M_0: 영(null)모형 (혹은 절편항 모형)
- Step 2: 각 $k = 1,2, \ldots, p$에 대해
  - k개의 예측변수를 가진 모형을 적합시킴
  - $\binom{p}{k}$개 모형 중에서 RSS를 가장 작게 하는(혹은 R^2를 가장 크게 하는) 모형을 선택 ⇒ M_k
- Step 3: CV 예측 MSE(혹은 오류율), C_p, AIC, BIC, 수정된 R^2를 써서 $M_0, M_1, \ldots, M_p$ 중에서 최적 모형을 선택

최적 부분집합 선택: 신용자료

- 예측변수는 10개
- 민족(ethnicity)는 질적 변수: 범주 수 3개(가변수 2개 필요)
- 그림6.1에서 가로축(예측변수 개수) 값: 1-11
- 예측변수 3개 이상부터는 RSS(혹은 R^2)의 개선이 거의 없음

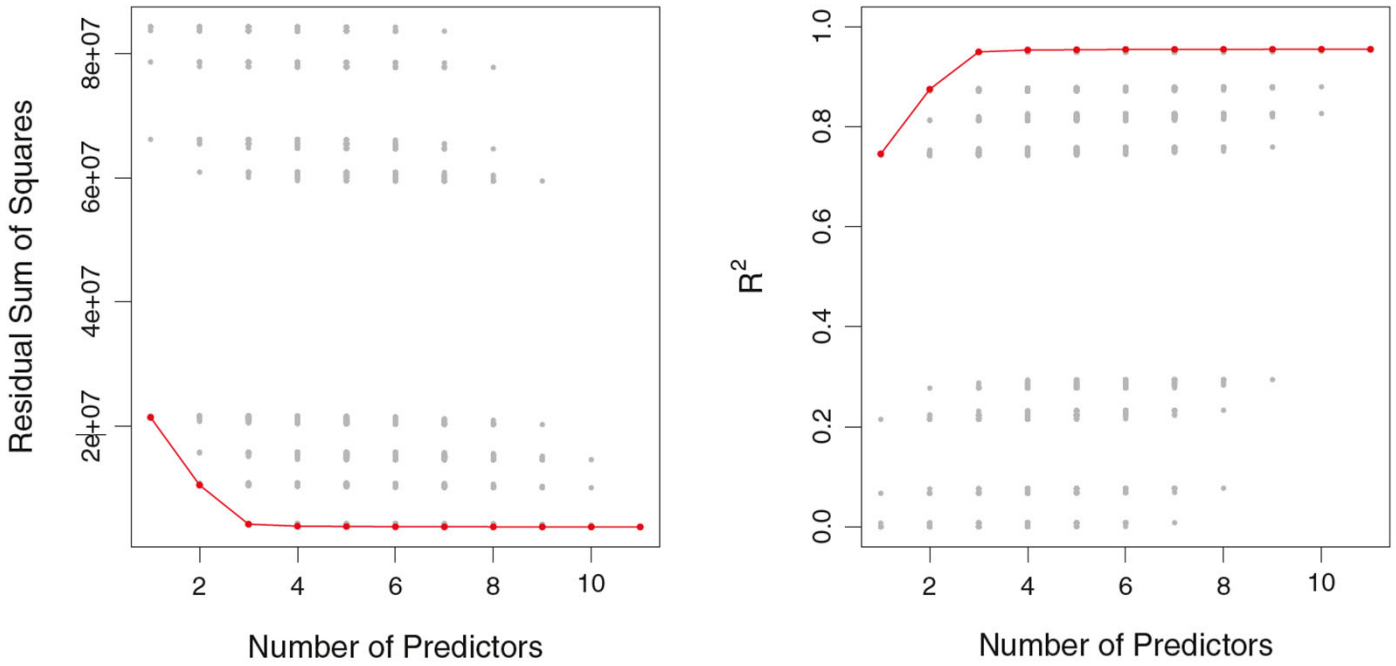


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

최적 부분집합 선택

- Remark 1: 로지스틱 회귀모형에도 적용 가능. 다만 RSS 대신에 deviance(이탈도)를 이용하고 deviance가 작은 모형을 선택
  - 잔차(residual) 이탈도: $2 \times \left(LL(\text{포화 모형}) - LL(\text{제안 모형}) \right)$,

$$LL = \log(\text{likelihood})$$

- Remark 2: p의 값이 크면 계산양이 폭증. 몇몇 모형은 계산하지 않는 지름길(shortcuts)이 요구됨. 즉 분기한정(branch-and-bound technique) 알고리즘이 요구됨
- Remark 3: 탐색 모형이 많아 과대적합(overfitting)에 노출

단계적 선택(stepwise selection)

- 전진선택 방법(forward selection; FS)과 후진선택 방법 (backward selection; BS), 혼합 방법(hybrid approach)

전진선택: 알고리즘

- Step 1: M_0: 영(null)모형
- Step 2: 각 $k = 0,1, \dots, p-1$에 대해
  - 모형 M_k에 포함되지 않은 $(p-k)$개 예측변수 중에서 한 예측변수만 M_k에 추가하여 모형을 적합시킴
  - $(p-k)$개 모형 중에서 RSS를 가장 작게 하는(혹은 R^2를 가장 크게 하는) 모형을 선택 ⇒ M_{k+1}
- Step 3: CV 예측 MSE(혹은 오류율), C_p, AIC, BIC, 수정된 R^2를 써서 $M_0, M_1, \dots, M_p$ 중에서 최적 모형을 선택

전진선택

- Remark 1: 적합시키는 모형 개수 $= 1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$
- Remark 2: 2^p개의 부분집합 중에서 반드시 최적 모형을 선택한다고 할 수 없음
  - 신용자료: $M_1 - M_3$는 같지만 M_4는 서로 다름
- Remark 3: 고차원(high-dimensional)자료 즉 $n < p$일 때 FS 방법을 적용가능 하지만 $M_0, M_1, \ldots, M_{n-1}$까지만 선택 가능

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

후진선택: 알고리즘

- Step 1: M_p: 완전(full)모형. 즉 모든 예측변수를 포함한 모형
- Step 2: 각 $k = p, p+1, \ldots, 1$에 대해
  - 모형 M_k에 포함된 k개의 예측변수 중에서 한 예측변수만 제거하여 모형을 적합시킴
  - k개 모형 중에서 RSS를 가장 작게 하는(혹은 R^2를 가장 크게 하는) 모형을 선택 ⇒ M_{k-1}
- Step 3: CV 예측 MSE(혹은 오류율), C_p, AIC, BIC, 수정된 R^2를 써서 $M_0, M_1, \ldots, M_p$ 중에서 최적 모형을 선택

후진선택

- Remark 1: FS와 마찬가지로 적합시키는 모형 개수 $= 1 + \frac{p(p+1)}{2}$ 이고, 2^p개의 부분집합 중에서 반드시 최적 모형을 선택한다고 할 수 없음
- Remark 2: 고차원자료에 적용 불가능

혼합 방법

- FS 방법으로 진행하되 모형에 포함된 변수 중 더 이상 개선효과가 없는 변수는 제거
- 최적 모형에 더 가까운 모형을 선택할 수 있어

최적 모형 선택

- 예측변수 개수가 서로 다른 모형들 중에서 최적 모형을 선택할 때 RSS 혹은 R^2는 좋은 측도가 아님
- 테스트 오류율을 가장 작게 하는 모형을 최적 모형으로!
- 테스트 오류를 직접 추정하는 방법(5장)과 과대적합으로 인한 편의를 보정하는 간접적인 방법을 써서
  - 4 가지 대표적인 측도: C_p, AIC(Akaike information criterion), BIC(Bayesian information criterion), 수정된(adjusted) R^2

C_p, AIC

- 정의: $C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$
  - d: 예측변수 개수
  - $\hat{\sigma}^2$: 완전(full)모형으로부터 추정
- 정의: $\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$
- C_p(혹은 AIC)가 작을수록 최적 모형

C_p, AIC

- Remark: '$2d\hat{\sigma}^2$'는 벌점(penalty)에 해당. 모형에 포함되는 예측변수 개수가 많아질수록 RSS는 감소하지만 벌점은 점차 증가

BIC

- 정의: BIC $= \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + \log(n)\, d\hat{\sigma}^2\right)$
- Remark: 관찰 개체수가 많을수록 벌점이 커짐$(\log n > 2, n > 7)$

수정된 R^2

- 정의: 수정된 $R^2 = 1 - \frac{\frac{\mathrm{RSS}}{n-d-1}}{\frac{\mathrm{TSS}}{n-1}}$
- 수정된 R^2가 클수록, $\frac{\mathrm{RSS}}{n-d-1}$가 작을수록 최적 모형

수정된 R^2

- Remark: 모형이 제대로 적합되었다면 그 모형에 noise 변수 하나를 추가하면 RSS는 매우 조금 감소하지만 d가 증가하므로 궁극적으로 $\frac{\mathrm{RSS}}{n-d-1}$이 커지게 됨. 따라서 '수정된 R^2'는 감소

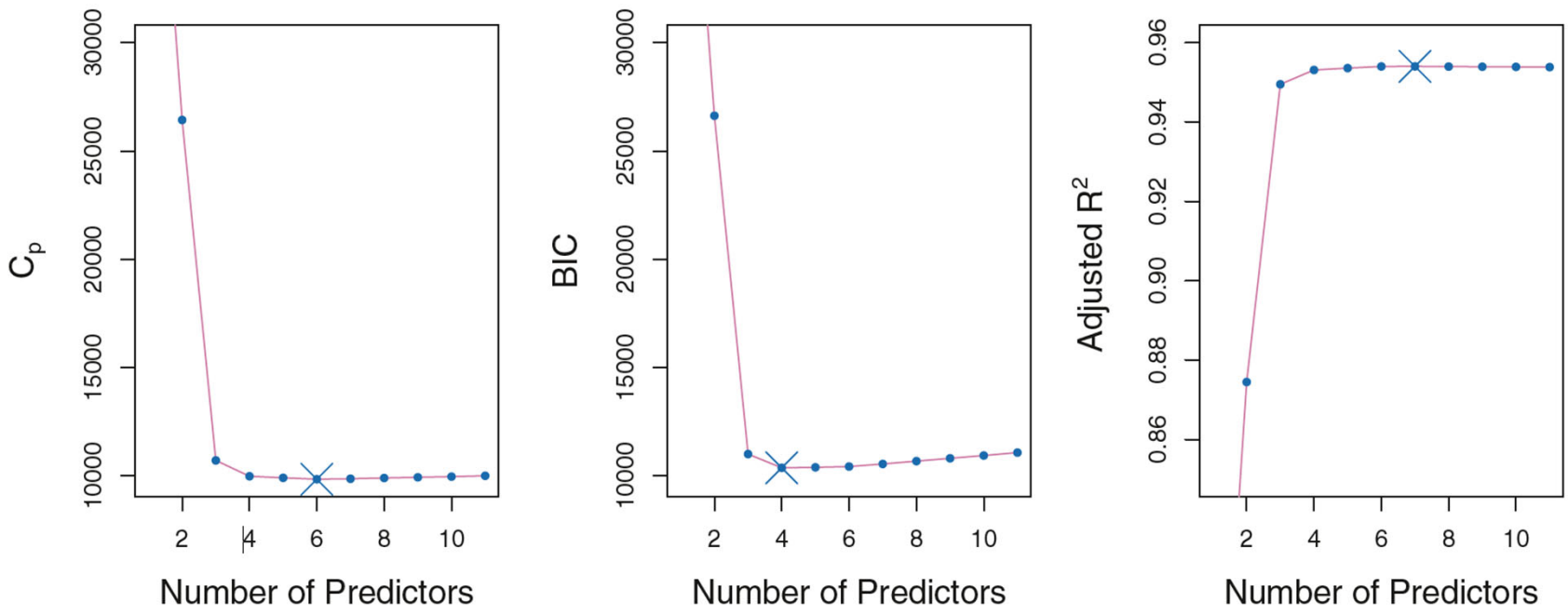


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

검증 & 교차검증

- 테스트 오류율을 직접 계산할 수 있고,
- 가정이 적으며,
- 모형선택문제에 폭넓게 사용 가능
- 계산양이 많지만 매력적인 방법

최종모형 선택 기준: 신용자료

- BIC, 검증, 교차검증 비교
- 검증: 3:1로 나눔
- 교차검증: 10-폴드

최종모형 선택 기준: 신용자료

- one-standard-error-rule: 검증, 교차검증에서는 관찰 개체의 분할(split)에 따라 최소 테스트 MSE가 변할 수 있음
  - 테스트 MSE의 표준오차를 계산
  - 최소의 테스트 MSE를 가진 테스트 MSE의 1-표준오차 내에 포함되는 테스트 MSE를 가진 모든 모형을 선택
  - 가장 예측변수가 작은 모형을 최종모형으로 선택하는 규칙 ⇒ 예측변수 개수 = 3인 모형을 선택

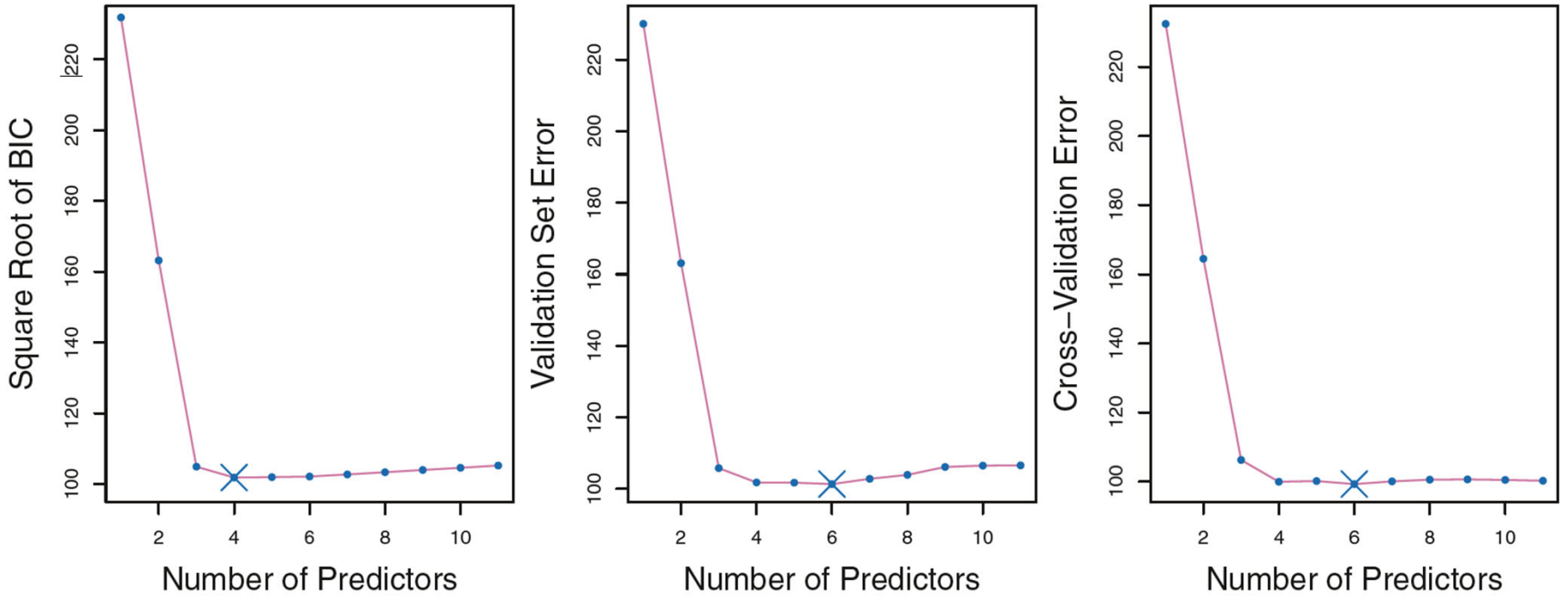


FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Thank you!

Move on to 06 Linear model section and regularization: Part II