4.1 회귀직선의 오차

1) 실제값과 추정치의 차이
2) 상관계수를 이용한 RMSE의 계산
3) 잔 차 도
4) 세 로 띠
5) 세로띠 별 분포를 정규분포로 근사시키기

1. 실제값과 추정치의 차이

제곱근-평균-제곱 오차 (RMSE)

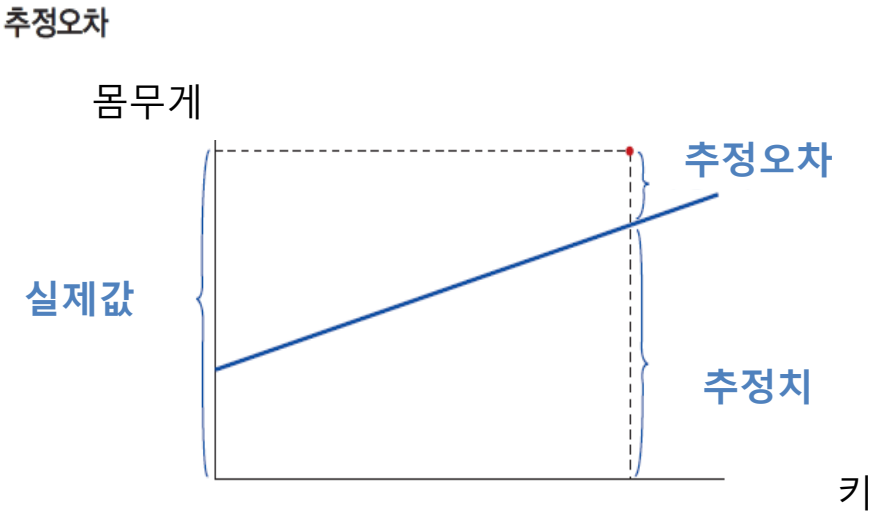
- 실제 값과 예측치의 차이가 어느 정도 될지 알려줌

- 추정의 표준오차(standard error of estimate) 또는 회귀의 표준오차 (standard error of regression)라고도 불림

1. 실제값과 추정치의 차이

추정오차 1

- 추정오차
  - 실제 몸무게 – 예측된 몸무게
  - 일반적으로 잔차(residual)라고 부른다.
  - 전반적인 크기는 제곱근-평균-제곱(RMS) 방식으로 측정한다.



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 144

주: 추정오차는 실제값으로부터 회귀직선까지의 수직거리와 같다.

1. 실제값과 추정치의 차이

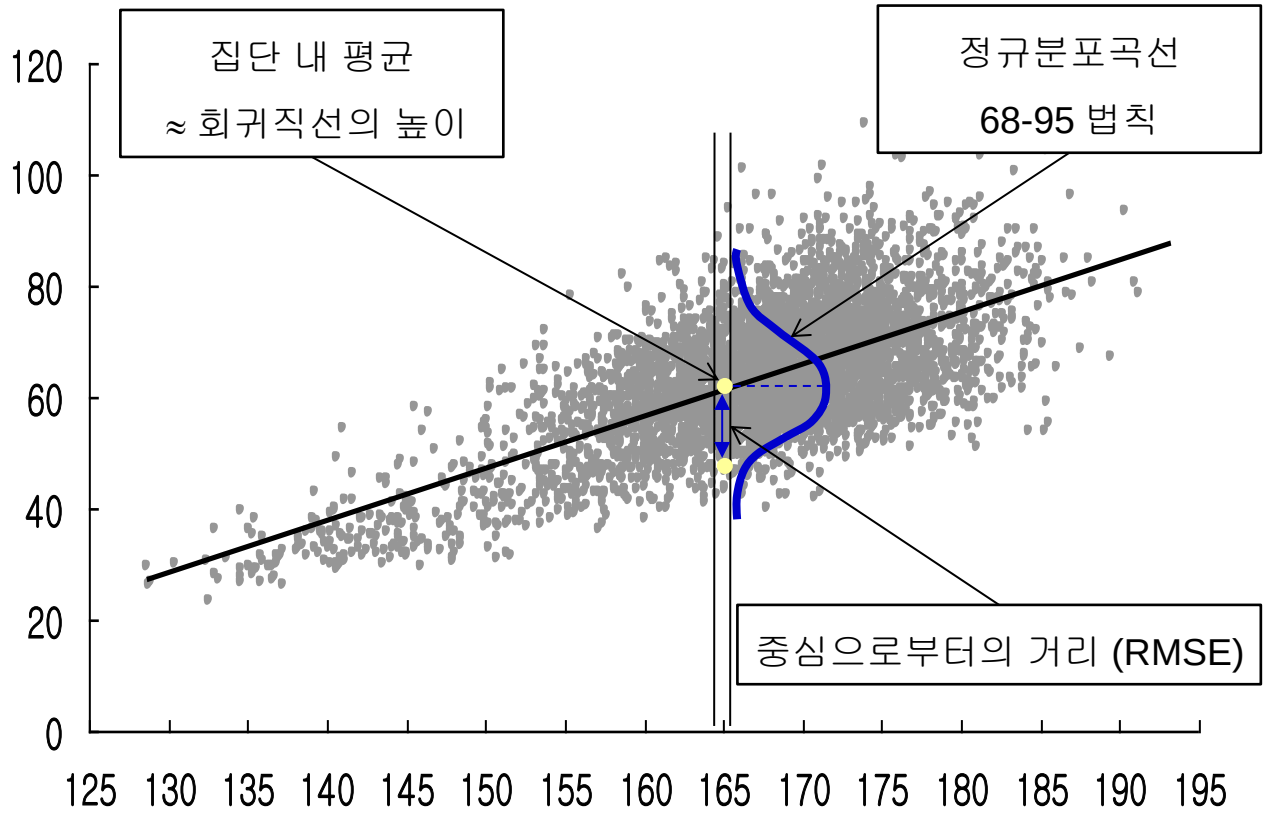
RMSE 구하는 방법

$$\mathrm{RMSE}=\sqrt{\frac{(1\text{번째 오차})^2+(2\text{번째 오차})^2+\cdots+(1,503\text{번째 오차})^2}{1,503-2}}=9.3\mathrm{kg}$$

- RMSE 구하는 방법
  - 산포도에서 전형적인 점(typical point)은 회귀직선으로부터 위 또는 아래로 9.3kg 정도 떨어져 있다. 실제 몸무게는 추정된 몸무게와 약 9.3kg 정도 다르다.
  - 분모에 표본크기가 아닌 자유도가 사용되었다.
  - 자유도=1,503-2=표본크기-2
  - 추정오차 계산의 기준은 회귀직선인데 이는 절편과 기울기의 두 추정치에 의해 결정되므로 자유도는 2만큼 감소.

1. 실제값과 추정치의 차이

회귀직선과 RMSE



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 144

1. 실제값과 추정치의 차이

회귀직선과 RMSE

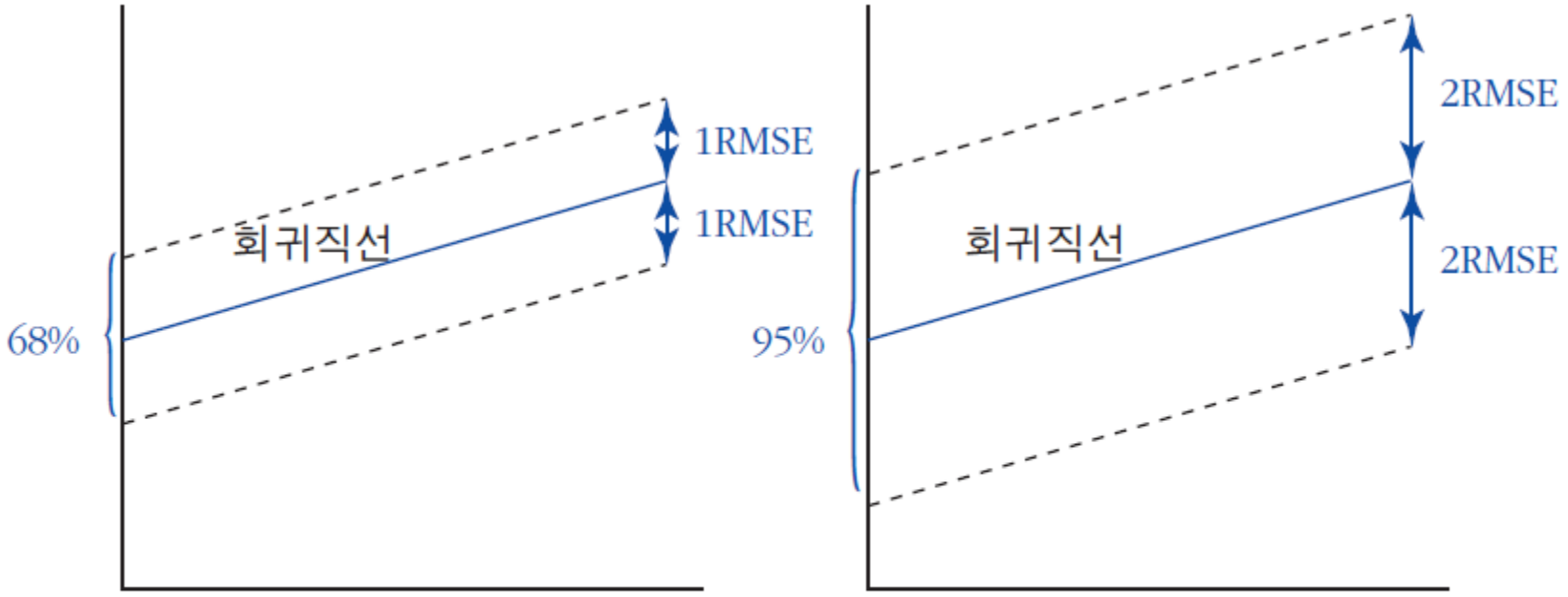
- 회귀직선은 x값에 따라 분류된 부분집단 별로 자료의 중심을 알려준다.

- RMSE는 개별 관측치가 그가 속한 준거집단의 평균으로부터 떨어진 정도를 대략적으로 알려준다.

- 회귀직선과 RMSE를 알면 평균과 표준편차를 알 때처럼 68-95 법칙을 활용해 볼 수 있다.

1. 실제값과 추정치의 차이

회귀직선, RMSE, 68-95법칙



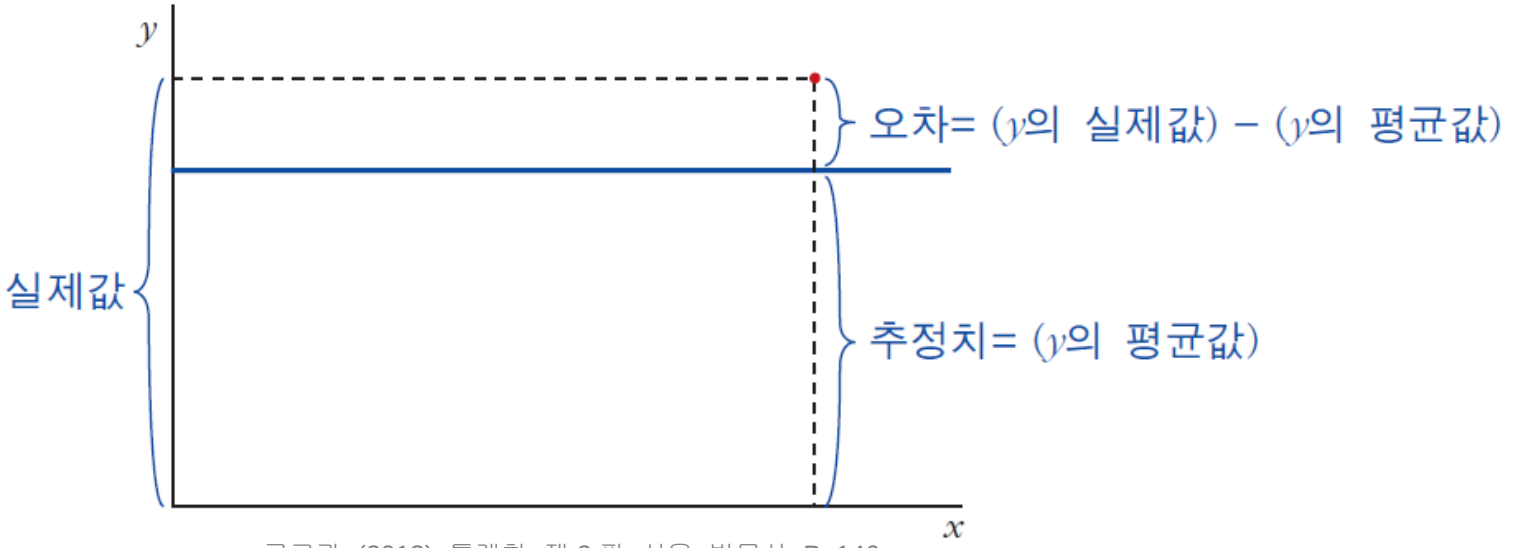
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 146

주: 산포도상의 점들 중 약 68%가 회귀직선으로부터 ±1RMSE만큼 떨어져 있는 두 직선 사이에 있다. 약 95%의 점들은 ±2RMSE만큼 떨어져 있는 두 직선 사이에 존재한다.

1. 실제값과 추정치의 차이

초보적인 추정방법의 RMSE

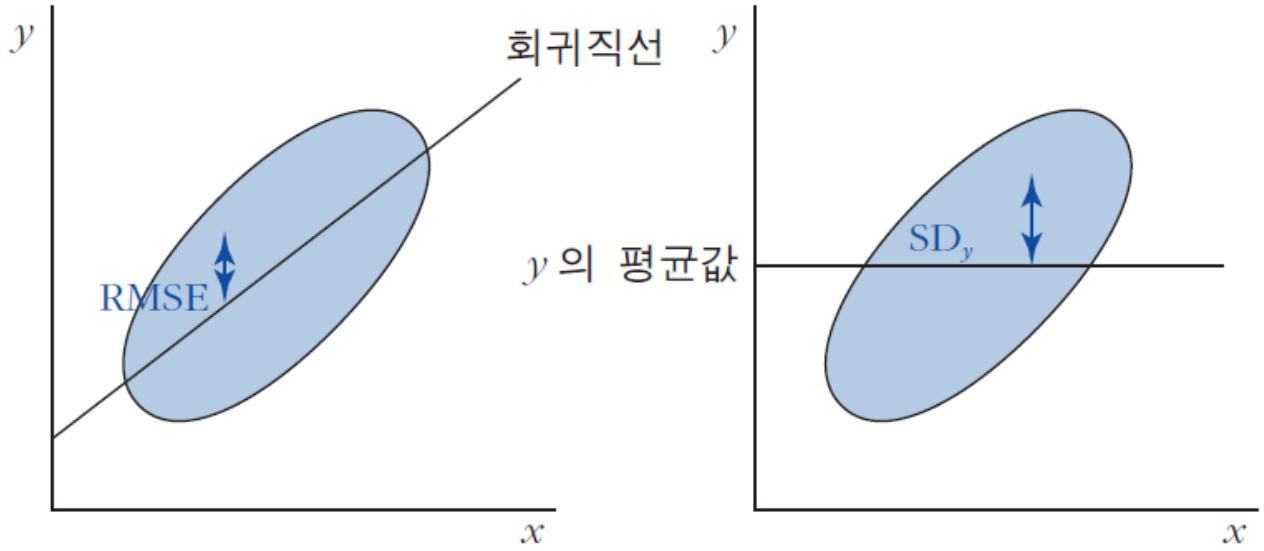
- 초보적인 추정방법: x값은 무시한 채 y값의 전반적인 평균으로 개별 y값을 추정 → (x, y) 그래프 상에서 y값 추정치들이 수평선을 이룬다.
- 초보적인 추정방법의 RMSE는 y의 표준편차(SD_y)가 된다.



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 146

2. 상관계수를 이용한 RMSE의 계산

회귀직선의 RMSE와 y의 표준편차



류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 147

- 일반적으로 '회귀직선의 RMSE < y의 표준편차.' 이는 수평선보다 회귀직선이 산포도상의 점들에 보다 가까이 위치하기 때문이다.
- 회귀직선의 RMSE는 대략 $\sqrt{1-r^2} \times SD_y$ 와 같다. (단, r은 x와 y의 상관계수)

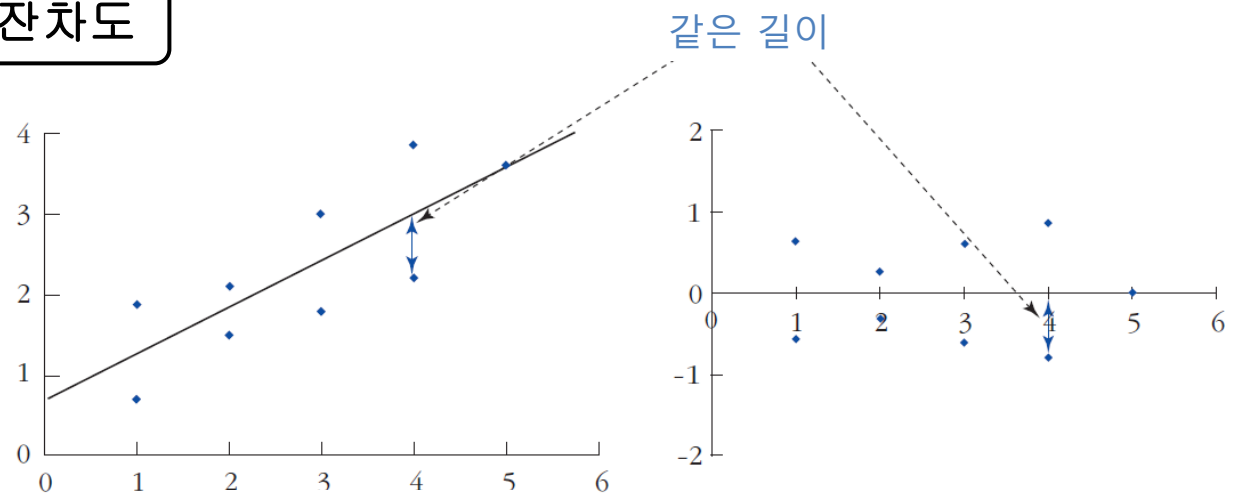
2. 상관계수를 이용한 RMSE의 계산

상관계수와 회귀직선의 RMSE

- $r = 1$ 경우
  - 산포도상의 모든 점들이 하나의 우상향하는 직선 위에 놓임
  - 추정오차는 모두 0. RMSE=0.
- $r = -1$ 경우
  - 산포도상의 모든 점들이 하나의 우하향하는 직선 위에 놓임
  - 추정오차는 모두 0. RMSE=0.
- $r = 0$ 경우
  - 두 변수 x와 y간에 선형관계가 전혀 없음
  - 회귀직선은 x값으로 부터 y값을 추정하는 데 전혀 도움이 안됨
  - RMSE는 SD_y 와 대략 같은 값을 갖게 된다.

3. 잔차도

일반적인 잔차도



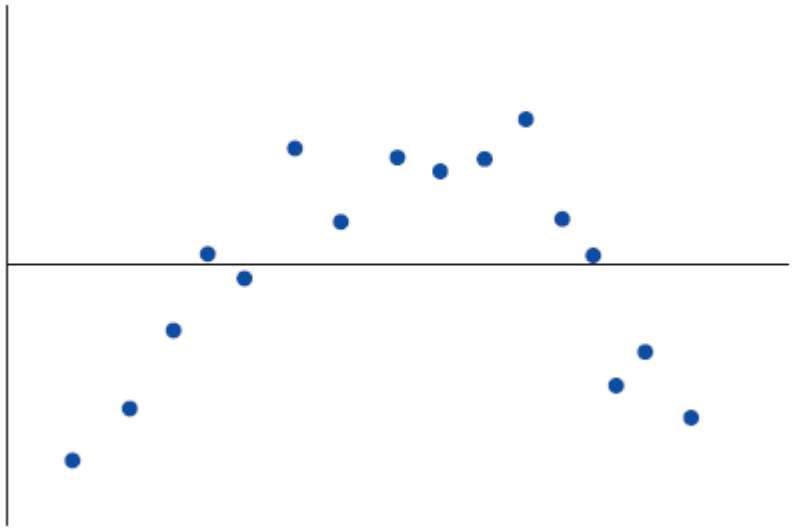
산포도와 회귀직선 잔차도

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 149

- 잔차들의 합도 0이고 잔차들의 평균도 0
- 잔차도 상의 점들은 우상향하거나 우하향하는 등의 체계적인 선형패턴 (linear pattern)을 보이지 않음. 산포도 상에서 관찰된 두 변수간 선형패턴은 이미 회귀직선에 흡수되어 버렸기 때문임

3. 잔차도

비선형의 패턴을 보이는 잔차도



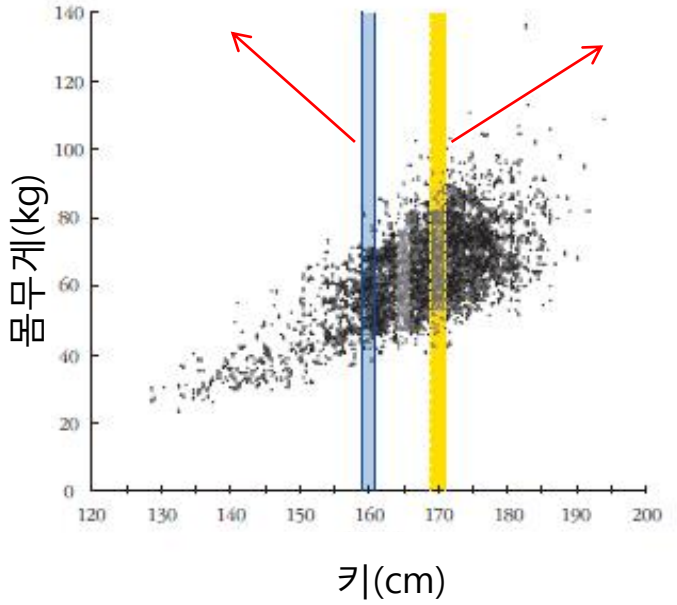
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 150

- 잔차도가 어떤 체계적인 패턴을 보이는 경우 회귀분석 모형에 무언가 문제가 있다고 보아야 함
- 잔차도에 남아 있는 뚜렷한 비선형의 패턴은 직선의 회귀분석 모형이 체계적인 비선형의 관계를 포착하지 못하고 누락시켰다는 점을 시사함

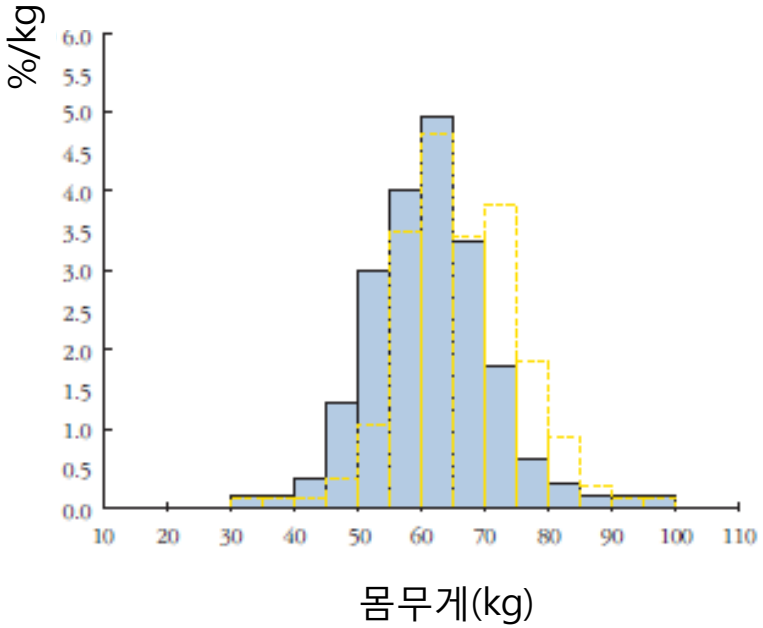
4. 세로띠

산포도와 세로띠 내의 히스토그램

키가 대략 **160 cm**인 사람들의 집



키가 대략 **170 cm**인 사람들의 집합



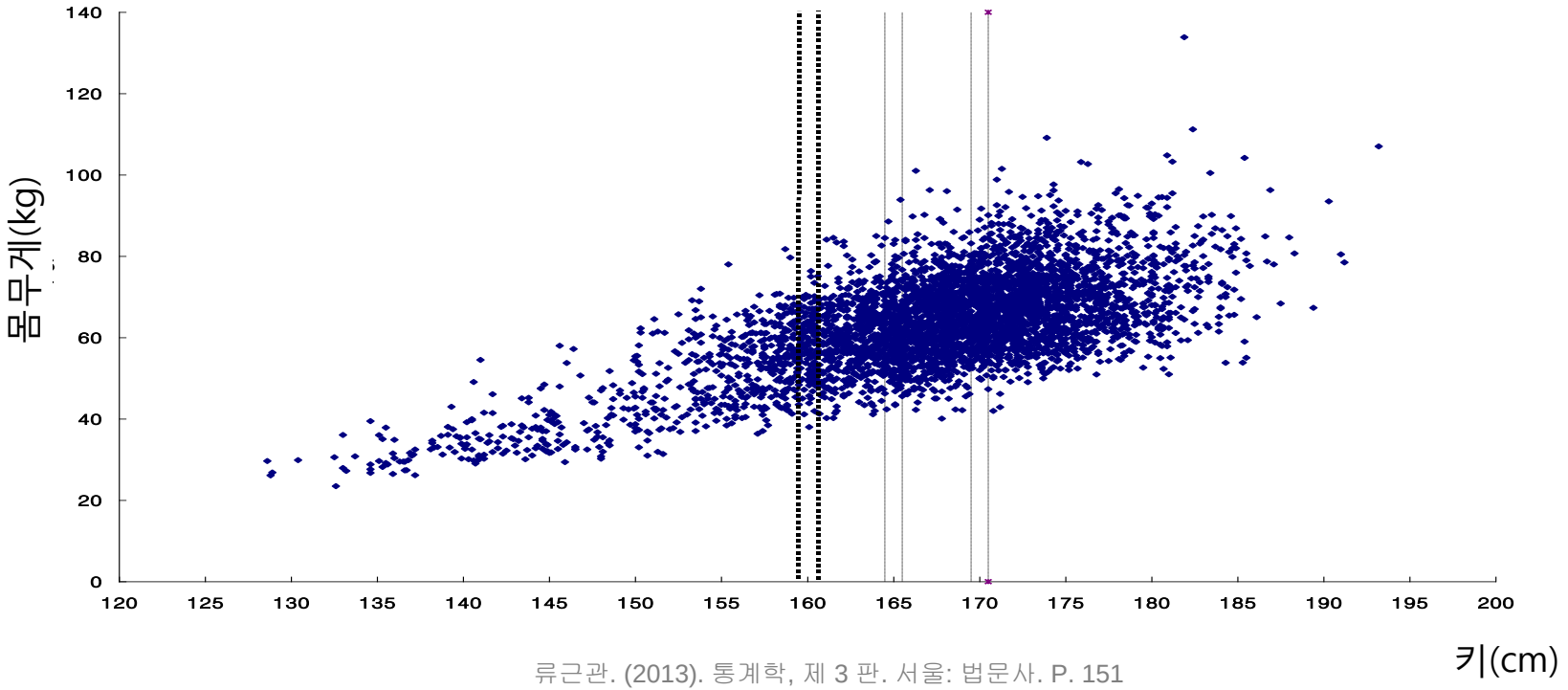
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 151

- 좌측의 두 세로띠에 해당되는 우측의 두 히스토그램을 비교해보면, 중심은 다르지만 퍼진 정도는 거의 같다.

4. 세로띠

등분산성과 이분산성

- 등분산성 (等分散性, homoscedasticity)



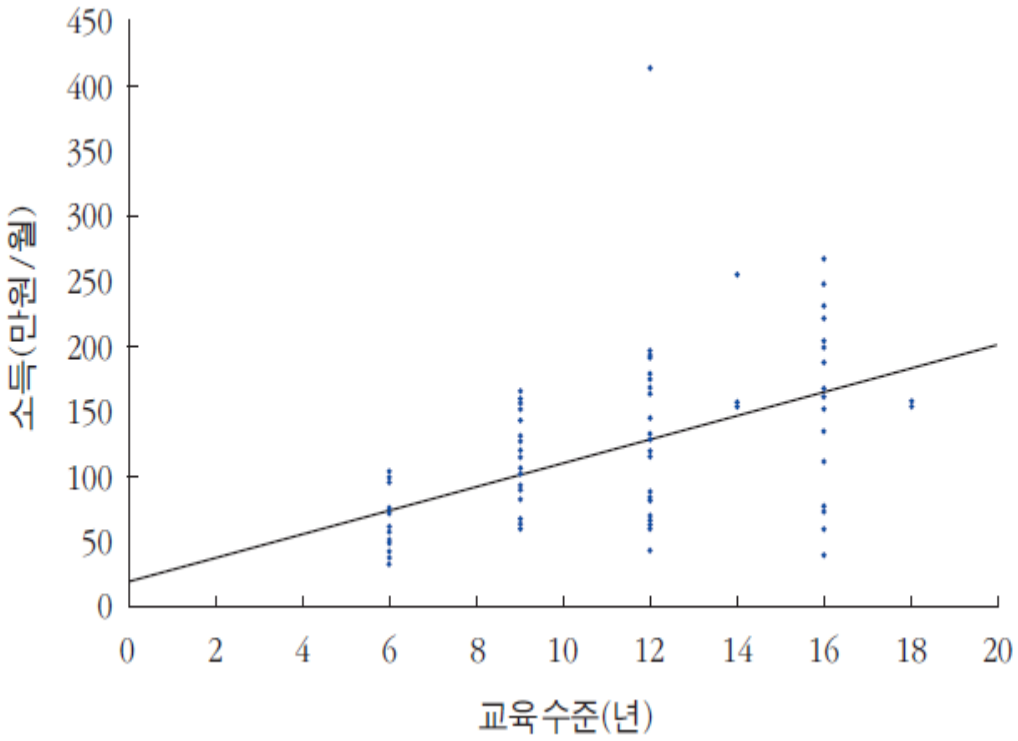
류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 151

- 회귀직선을 중심으로 점들이 위 아래로 퍼진 정도가 세로띠 별로 같음

4. 세로띠

등분산성과 이분산성

- 이분산성(異分散性, heteroscedasticity)



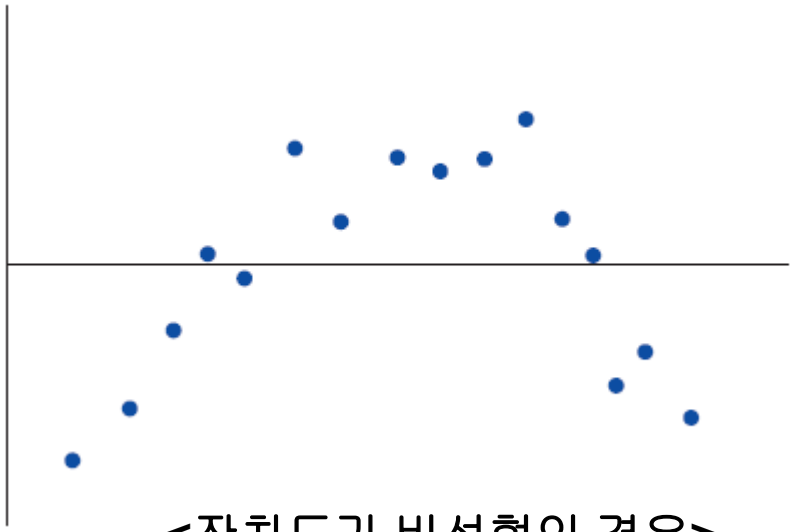
1) 산포도가 이분산성을 보일 때, 실제의 y값이 회귀직선에 의한 y값 추정치로부터 벗어나는 정도는 x값 별로 즉 세로띠 별로 달라짐.

2) 이분산성 존재 시, 회귀직선의 RMSE는 서로 다른 x값에 대응하는 추정오차들의 전반적인 크기를 나타낼 뿐이다.

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 152

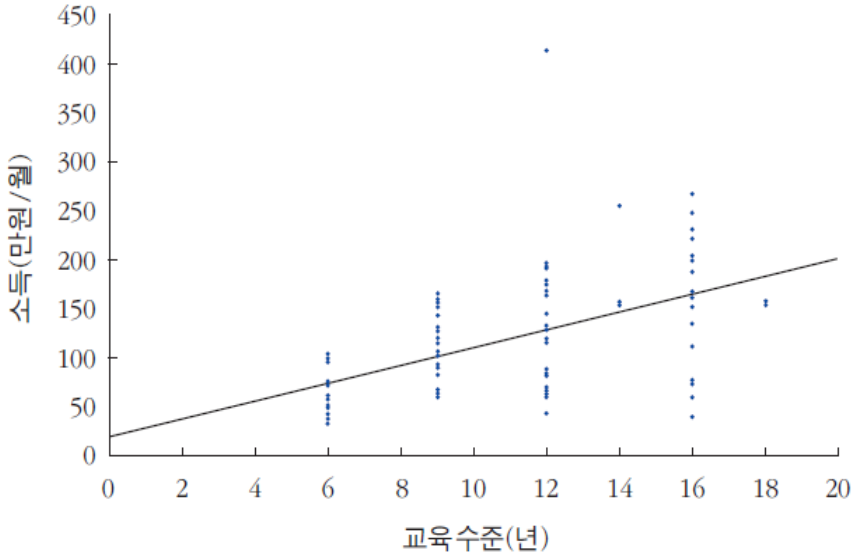
5. 세로띠 별 분포를 정규분포로 근사시키기

근사시킬 수 없는 경우



<잔차도가 비선형인 경우>

회귀직선으로 구한 y값 추정치가 부적절



< 이분산성을 띠는 경우>

공통의 RMSE가 부적절

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 150

류근관. (2013). 통계학, 제 3 판. 서울: 법문사. P. 152

5. 세로띠 별 분포를 정규분포로 근사시키기

정규분포 근사 예제

예: 2013학년도 1학기 한 대학에서 통계학을 수강한 학생들의 중간고사 점수(0-50)와 기말고사 점수(0-100)를 조사하였다.

중간고사 평균 = 27.9　　중간고사 표준편차 = 8.5

기말고사 평균 = 56.4　　기말고사 표준편차 = 13.8

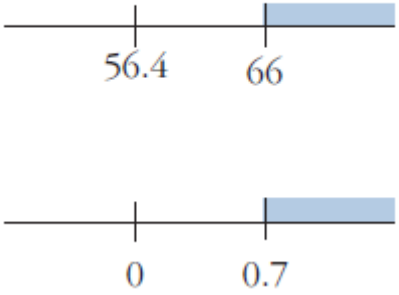
상관계수 = 0.49

- 산포도는 타원형이다.
  1) 기말고사 성적이 66점 이상인 학생은 전체 몇 % 정도인가?
  2) 중간고사 점수가 33점인 학생들 중에서 기말고사 점수가 66점 이상인 학생은 전체의 몇 %쯤 되는가?

5. 세로띠 별 분포를 정규분포로 근사시키기

정규분포 근사 예제

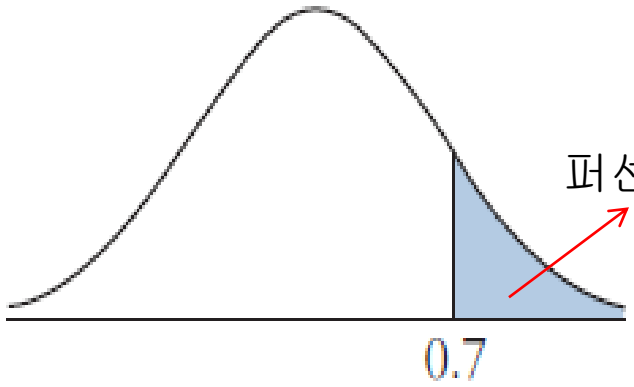
1) 중간고사 관련 통계치나 상관계수 정보는 필요하지도 않다.



$$\frac{66-56.4}{13.8}=0.7$$

← y의 평균

y의 표준편차



퍼센트 = 음영부분 ≈ 24%

5. 세로띠 별 분포를 정규분포로 근사시키기

정규분포 근사 예제

2) 회귀분석으로 "새로운 평균"(회귀직선으로 구한 y값 추정치)과 "새로운 표준편차"(RMSE)를 구해 문제에 답한다.

1) 중간고사 점수가 평균보다 0.6 SD_x 만큼 높다.
2) 상관계수는 0.49이다. 0.49×0.6 = 0.3
3) 기말고사 점수는 0.3 SD_y = 4.1점만큼 높다.
4) "새로운 평균"은 56.4 + 4.1 = 60.5점이 된다.
5) "새로운 표준편차"는 다음 식을 통해 구한다.

$$\sqrt{1-r^2} \times SD_y = \sqrt{1-0.49^2} \times 13.8 = 12$$