Finance Analytics
Chapter3. Linear Regression Model
Part 6. Discussion

권태연

한국외대 국제금융학과

표준오차

표준편차 σ 추정, 표준오차 $se(b_k)$추정

- Recall : 가정 4.
  - 가정 4. X값이 주어져 있을 때, 오차항의 분산은 σ^2로 모든 개체 i에 대해 동일하다.= 등분산(homoscedastoc)가정.
$$var(u_i|X) = \sigma^2$$
  - 가정 4'. X값이 주어져 있을 때, Y의 분산은 σ^2로 모든 개체 i에 대해 동일하다.= 등분산(homoscedastoc)가정.
$$var(Y_i|X) = \sigma^2$$
- 분산 σ^2의 추정은 잔차(residual)의 분산을 이용한다.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-p}$$

  이때 p는 추정해야하는 B의 갯수, X의 갯수+1
- 표준오차 $se(b_k)$는 독립변수간에 상관관계가 없다면 (가정 6)

$$\hat{s}.e(b_k)^2 = \frac{\hat{\sigma}^2}{\sum(x_{ik} - \bar{x}_k)^2}$$

- 시간당 임금함수 예제에서 분산, 회귀계수의 표준오차 확인 및 해석

가정7. 모형설정오류

모형설정오류 -핵심변수의 누락

omitted variable bias

- Example: Meditation and Aging (Noetic Sciences Review, Summer, 1993, p28)
- 설명변수: 명상수련 여부
- 반응변수: 노화 정도
- Lurking Variable (잠복변수): These have an important effect on the relationship among the variables in a study, but are not included in the study.
- Lurking variable:

모형설정오류 -핵심변수의 누락

- 다중회귀분석계수와 단순회귀분석 계수가 차이나는 이유?
  - 다중회귀모형에서 회귀계수 (slope)의 의미를 생각해보자.
  - 가정 6에 대하여 생각해보자.
- Control, Adjust Lurking Variables.
- Lurking Variable (잠복변수): These have an important effect on the relationship among the variables in a study, but are not included in the study.

모형설정오류- 불필요한 설명변수

- 불필요한 설명변수의 포함
  - Causation과 Correlation은 다름!
    - Correlation : 인과관계의 방향성 없음!
    - Regression : 인과관계의 방향성 있음!
    - Statistics can lie!!!!
    - Example 1 : 아이스크림 판매량과 에어컨 판매량
    - Example 2 : 기온과 아이스크림 판매량
    - Example 3 : 맥주소비량과 혈중알콜농도
    - Example 4 : 아버지의 키와 아들의 키
    - Example 5 : 흡연과 암
  - 다중공선성 문제가 발생할 수 있음

모형설정오류- 과소적합(underfitting), 과적합(overfitting)

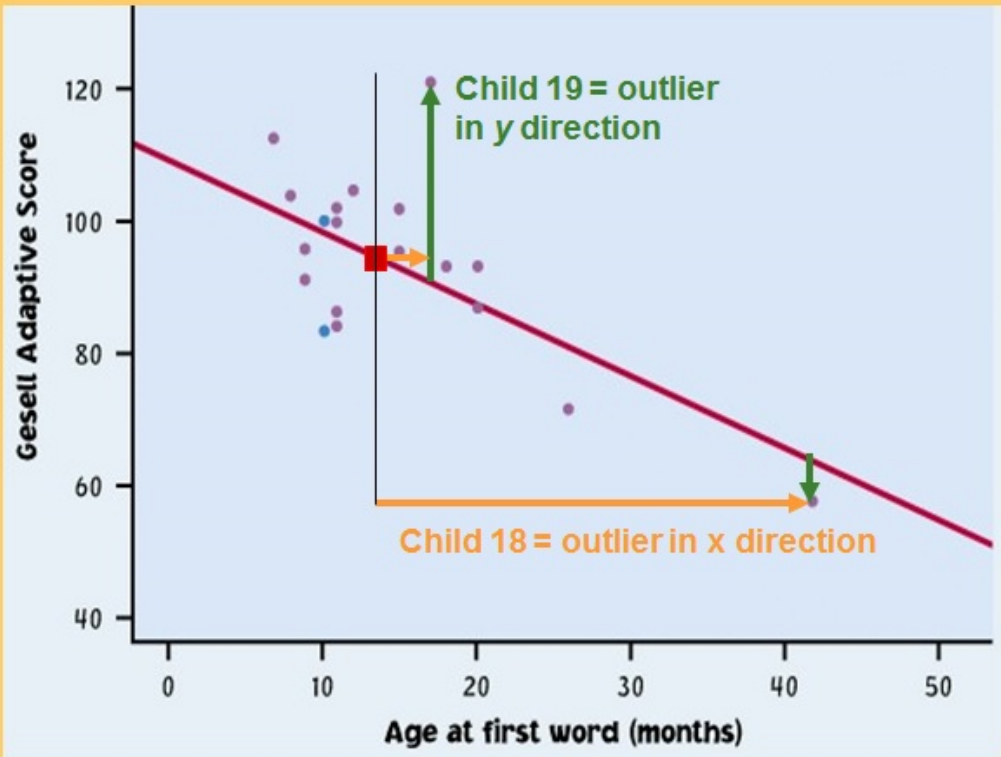
- 과소적합(underfitting) -핵심변수의 누락
  R^2가 너무 작아서 생기는 문제..
- 과적합(overfitting) - 불필요한 설명변수의 포함
  R^2가 너무 커서 생기는 문제..
- 두 경우 모두 예측에 사용하기 힘들다.

모형설정오류- 특이치 (Outlier)에 대한 고려

- Outlier: observation that lies outside the overall pattern of observations.
- Influential individual: observation that markedly changes the regression if removed. This is often an outlier on the x-axis.

모형설정오류- 특이치 (Outlier)에 대한 고려

Example: Does age at first word predict the score of later mental ability test in children?

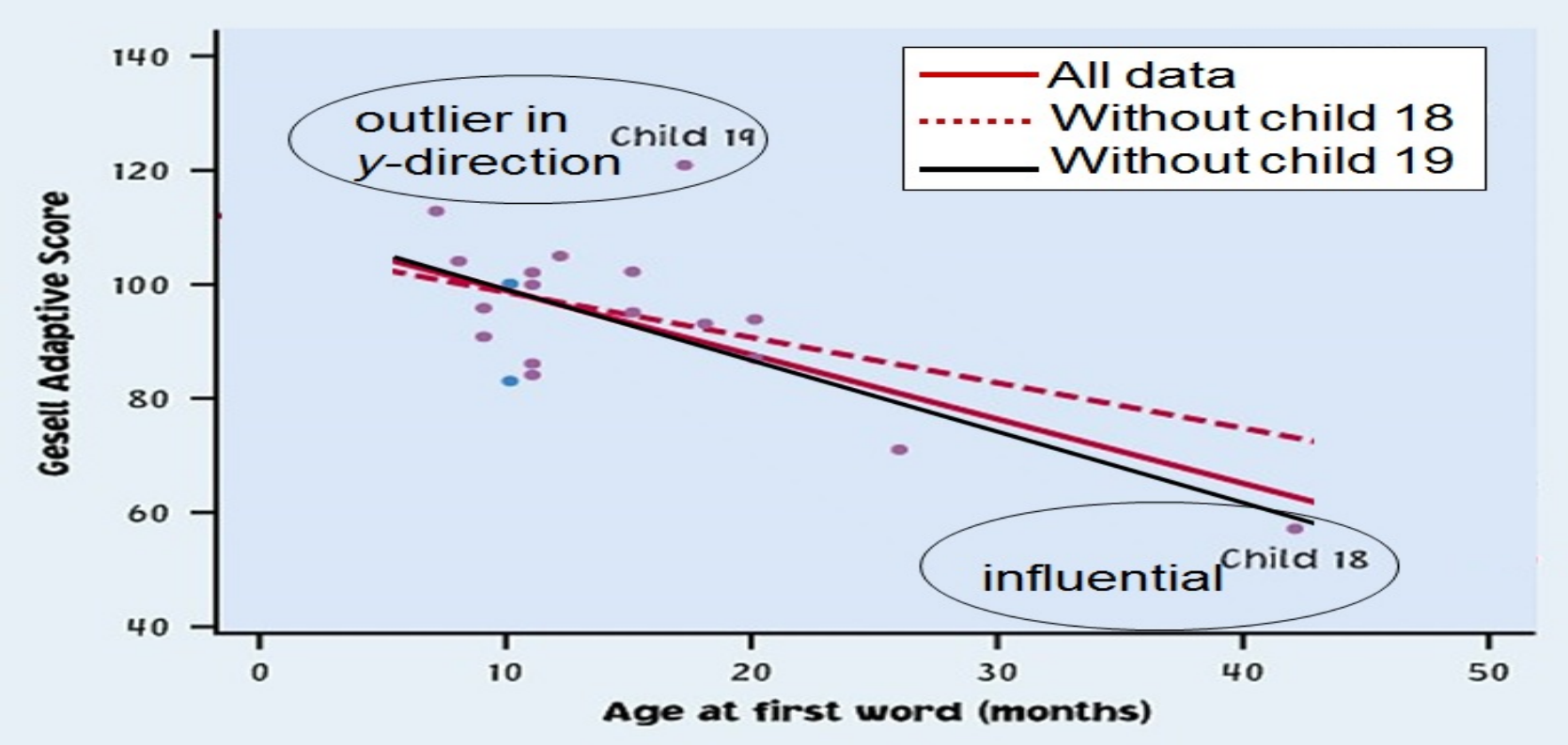


Child 19 is an outlier of the relationship.

Child 18 is only an outlier in the *x* direction and thus might be an influential point.

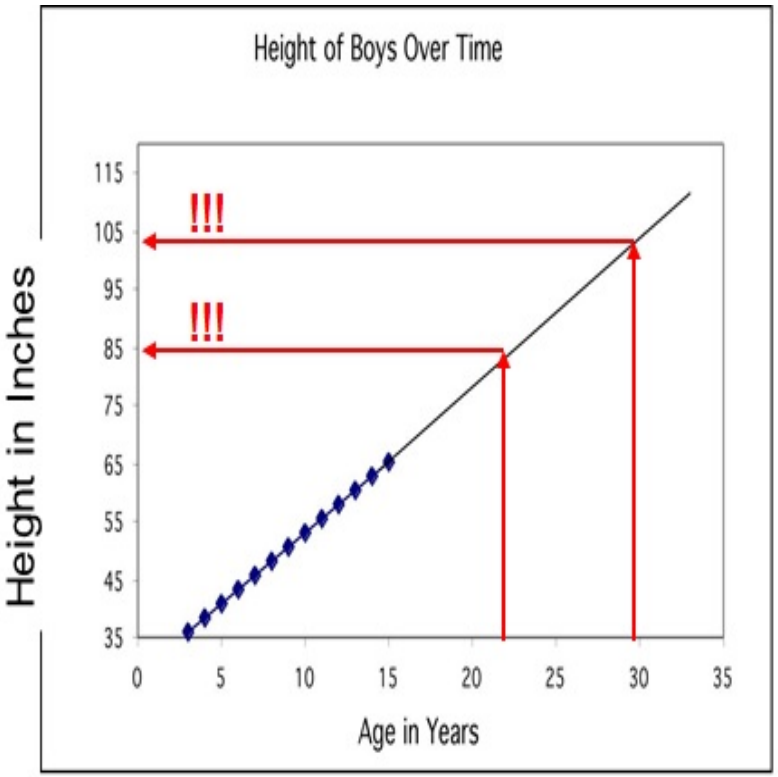
모형설정오류- 특이치 (Outlier)에 대한 고려

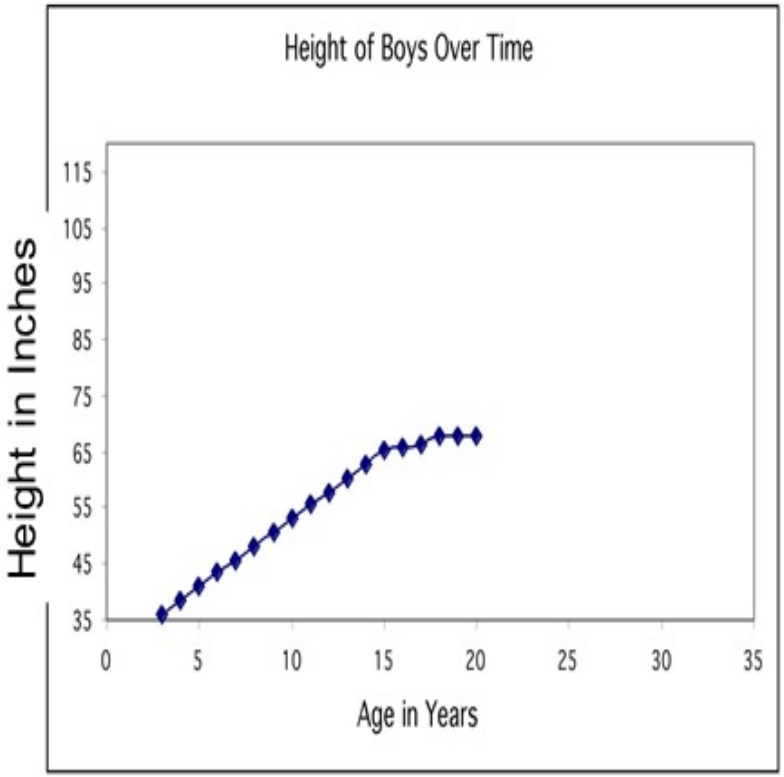
- A regression line can be dramatically affected by an influential point



모형설정오류- Extrapolation

- Extrapolation is the use of a regression line for predictions outside the range of x values used to obtain the line.
- This can be a very stupid thing to do, as seen here.



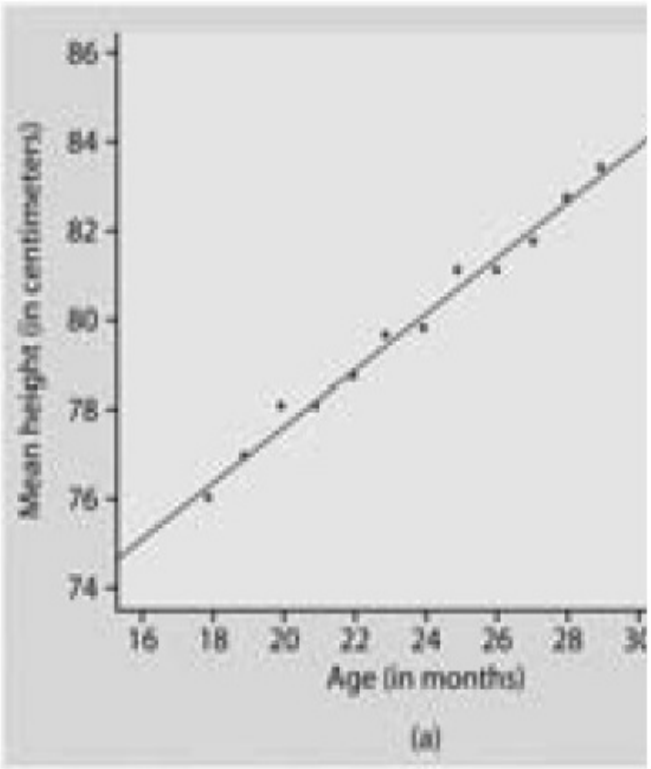


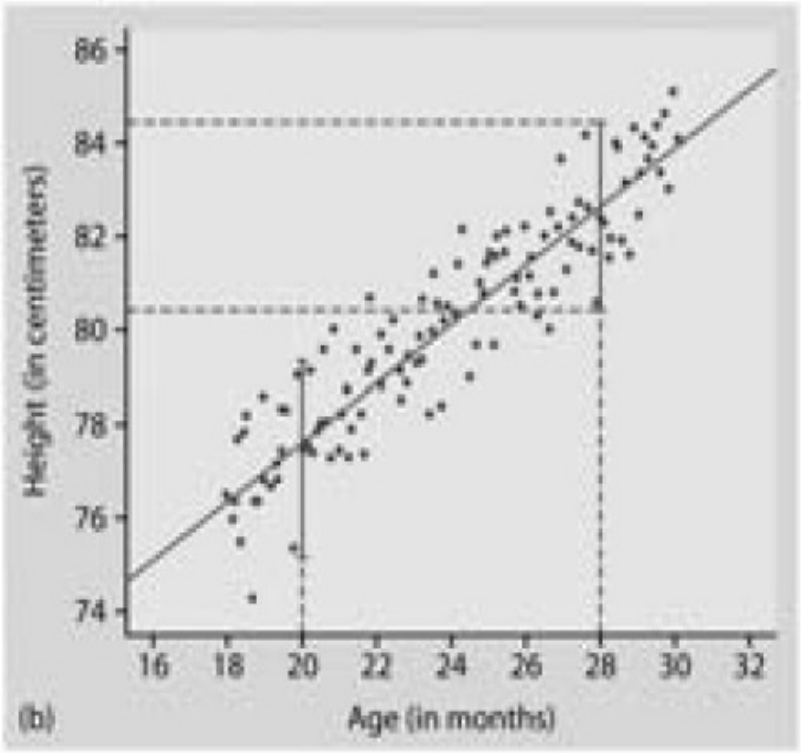
모형설정오류- Aggregation

- Aggregation : Association based on average data
- Problem: A scatter plot of just the average hides much of the variability in the data
- In general, regression with aggregate data overstates the strength of the association (larger r^2)

Aggregation: Age and Height of children

Aggregate: $r^2 = 0.989$ **Raw data: $r^2 = 0.849$**





모형설정오류-함수 형태

- 선형관계에 있지 않은 변수간 선형모형을 적합함으로 생기는 오류
- 적절한 변수변환으로 해결
- 해석상의 이유로 변수변환을 실시하는 경우도 있음

모형설정오류-오차항의 확률분포

- 정규분포
- 잔차의 histogram, 잔차의 q-q plot
- 표본의 크기가 큰 경우에는 문제되지 않음
  $\because$ 중심극한정리

가정2. 확률적 설명변수 ?

확률적 설명변수, 연립방정식 -어려운 문제

- 독립변수를 확률변수가 아닌 상수로 정의하기 힘들때
- 종속변수의 갯수가 한개가 아니고 그들간에 상관관계 역시 무시하기 힘들 때