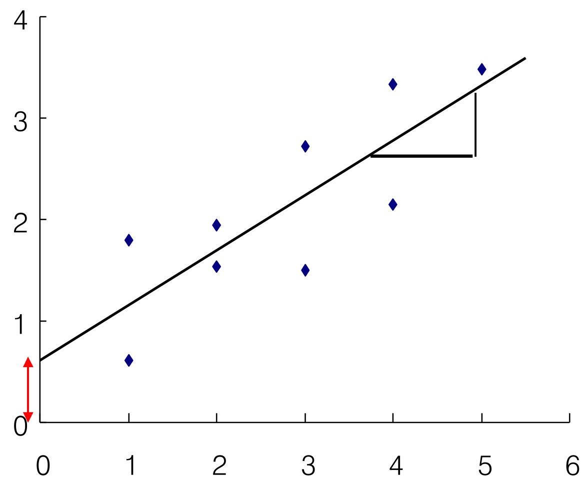
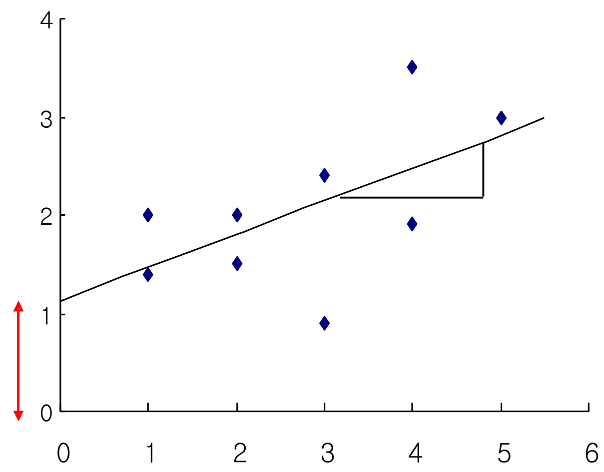
제 20 장 회귀분석과 유의성 검정

1. 개별 계수 추정량의 표준오차
2. 개별 계수에 대한 추론
3. 개별값 예측과 평균값 예측
4. Yellowstone Geyser “Old Faithful”
5. 회귀 잔차를 이용한 런검정
6. 런검정의 응용

1. 개별 계수 추정량의 표준오차

회귀계수 추정량의 표준오차





자료가 변하면 절편과 기울기가 모두 변한다.

이 때 얼마나 변화할지는 절편과 기울기 추정량의 표준오차를 보고 짐작할 수 있다.

1. 개별 계수 추정량의 표준오차

기울기와 절편 추정량의 표준오차

단순회귀분석에서 절편과 기울기 추정량의 표준오차

$$\mathrm{SE}(a)=\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\ ,\quad \mathrm{SE}(b)=\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\widehat{\mathrm{SE}}(a)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\ ,\quad \widehat{\mathrm{SE}}(b)=\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

σ: 회귀분석 오차의 표준편차

$\hat{\sigma}$: 회귀분석 오차의 표준편차에 대한 추정치

$$\hat{\sigma}=\sqrt{\sum_{i=1}^{n}(y_i-a-bx_i)^2/(n-2)}$$

앞에서는 이를 RMSE로 표기했었음

2. 개별 계수에 대한 추론

회귀분석 결과의 보고

단순회귀분석 결과의 보고

$$\hat{y} = \underset{(SE(a))}{a} + \underset{(SE(b))}{b}\, x$$

(단, 괄호 안은 표준오차)

관측치수=n, 결정계수=R^2, 추정의 표준오차=RMSE

2. 개별 계수에 대한 추론

기울기에 대한 추론

단순회귀분석모형의 기울기에 대해

귀무가설: β= 0. (즉, 기울기가 0 이다, x가 y를 설명 못한다.)

신뢰구간의 구축

- 기울기에 대한 95% 신뢰구간 $b \pm 2SE(b)$ 를 구해 β 를 포함하고 있는지 확인. 포함하지 않으면 5% 유의수준에서 귀무가설 기각

t-값의 계산

- 기울기 추정치의 t-값을 보고 판정
- $|t| \geq 2$ 이면 귀무가설 기각. 여기서 $t = \dfrac{b - \beta_0}{SE(b)}$

3. 개별값 예측과 평균값 예측

개별값 예측과 평균값 예측

광고비로 연간 1억원을 쓸 때

- 그 해 매출액이 얼마일까?   개별 y값의 예측

- 평균 매출액이 얼마일까?   평균 y값의 예측

- 예측치는 $\hat{y}_0 = a + bx_0$ 로 같고 그 예측의 표준오차만 서로 다르다.

3. 개별값 예측과 평균값 예측

개별값 예측의 표준오차 > 평균값 예측의 표준오차

개별값 예측의 표준오차에 대한 추정치

$$=\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$\widehat{\mathrm{SE}}(y_0-\hat{y}_0)$

$>$

평균값 예측의 표준오차에 대한 추정치

$$=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$\widehat{\mathrm{SE}}(\mathrm{E}y_0-\hat{y}_0)$

개별값을 예측하기가 평균값을 예측하기보다 더 어렵다.

3. 개별값 예측과 평균값 예측

개별값 예측의 예측구간 과 평균값 예측의 예측구간

단순회귀분석 모형에서 개별값의 95% 예측구간 :

$$\hat{y}_0 \pm 2\widehat{\mathrm{SE}}(y_0 - \hat{y}_0) = a + bx_0 \pm 2\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

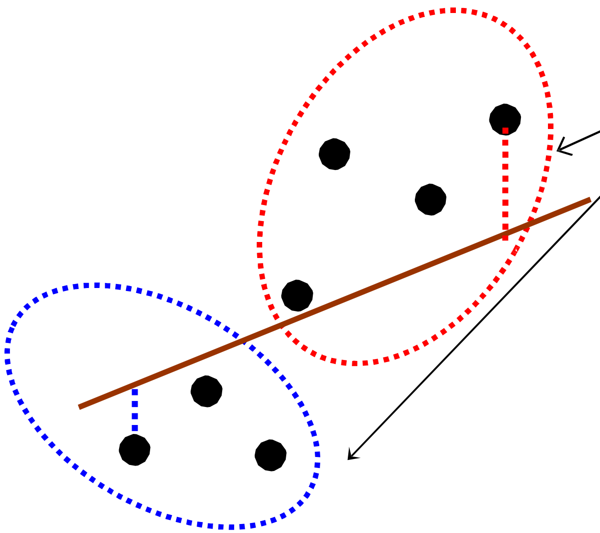
단순회귀분석 모형에서 평균값의 95% 예측구간 :

$$\hat{y}_0 \pm 2\widehat{\mathrm{SE}}(\mathrm{E}y_0 - \hat{y}_0) = a + bx_0 \pm 2\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

5. 회귀 잔차를 이용한 런검정

런과 계열상관

인접한 잔차간 상관관계



인접한 잔차간 부호를 비교:

같은 부호묶음을 하나의 런(run)으로 정의

+ → +
- → -

양의 계열상관을 띤다

+ → -
- → +

음의 계열상관을 띤다

5. 회귀 잔차를 이용한 런검정

런검정

++ ───── ++++++ ──── +++

- 런의 개수는 5개, 양(+)의 부호는 11개, 음(-)의 부호는 9개
- "+" 11개와 "–" 9개로 이루어진 자료에서 런개수의 하한 임계치는 6, 상한 임계치는 16
- "관측된 런의 개수 5 < 하한 임계치 6" 이므로 양의 계열상관 존재로 판정
- 이들 임계치는 많은 통계학 교재에 제시되어 있음. 컴퓨터만 있으면 이들 임계치를 각자 만들어 낼 수도 있음

(i) 11개의 +와 9개의 –가 든 상자에서 무작위 비복원추출로 총 20장의 전체 카드를 차례로 뽑아 런의 개수 셈

(ii) 위의 작업을 1,000회 반복

(iii) 이렇게 구한 총 1,000개의 런 개수를 오름차순으로 정리하면 50번째 값이 하한 임계치가 되고 950번째 값이 상한 임계치가 됨

5. 회귀 잔차를 이용한 런검정

런검정의 하한임계치

부록 [표-4]: 런검정의 하한임계치

#(+) \ #(-)	2	3	4	5	6	7	8	9	10	11
2										
3										
4										
5		2	2	3						
6		2	2	3	3					
7		2	2	3	3	3				
8		2	3	3	3	4	4			
9		2	3	3	4	4	5	5		
10		2	3	3	4	5	5	5	6	
11		2	3	4	4	5	5	6	6	7

5. 회귀 잔차를 이용한 런검정

런검정의 상한임계치

부록 [표-4]: 런검정의 상한임계치

#(+) \ #(-)	2	3	4	5	6	7	8	9	10	11
2										
3										
4										
5			9	10						
6			9	10	11					
7				11	12	13				
8				11	12	13	14			
9					13	14	14	15		
10					13	14	15	16	16	
11					13	14	15	16	17	17

6. 런검정의 응용

런검정의 응용 1

농구선수가 게임 중 슛에서 기복을 보이는지 검정

- 기복이 있으면 성공과 실패가 몰려있을 것임
- 관측된 자료: 20개 슛을 시도하여 9개 성공
- 런의 개수를 셈
- 9개의 성공(+)과 11개의 실패(−)에 해당되는 "런 개수의 하한 임계치"는 6
- 관측된 런의 개수가 하한 임계치 6 이하이면 양의 계열상관, 즉 기복이 있는 것으로 판정함

6. 런검정의 응용

런검정의 응용 2

주식가격의 움직임에 모멘텀이 있는지 검정

- 여기서 모멘텀이란 하나의 시계열이 보이는 방향지속성임
- 20거래일간의 주가자료 수집하여 오르거나 변화가 없으면 + , 내리면 $-$ 로 기록하여 +의 수, -의 수, 런의 개수를 센다.
- 런의 개수가 하한 임계치 이하이면 주식가격의 움직임에 모멘텀이 있다고 판단
- 모멘텀이 있는 것으로 판단되면 주가가 오르면 사고 내리면 파는 이른바 "모멘텀 거래전략"에 따라 주식투자를 해봄직하다. ("at your own risk, though!")