BIGDATA-PALETTE

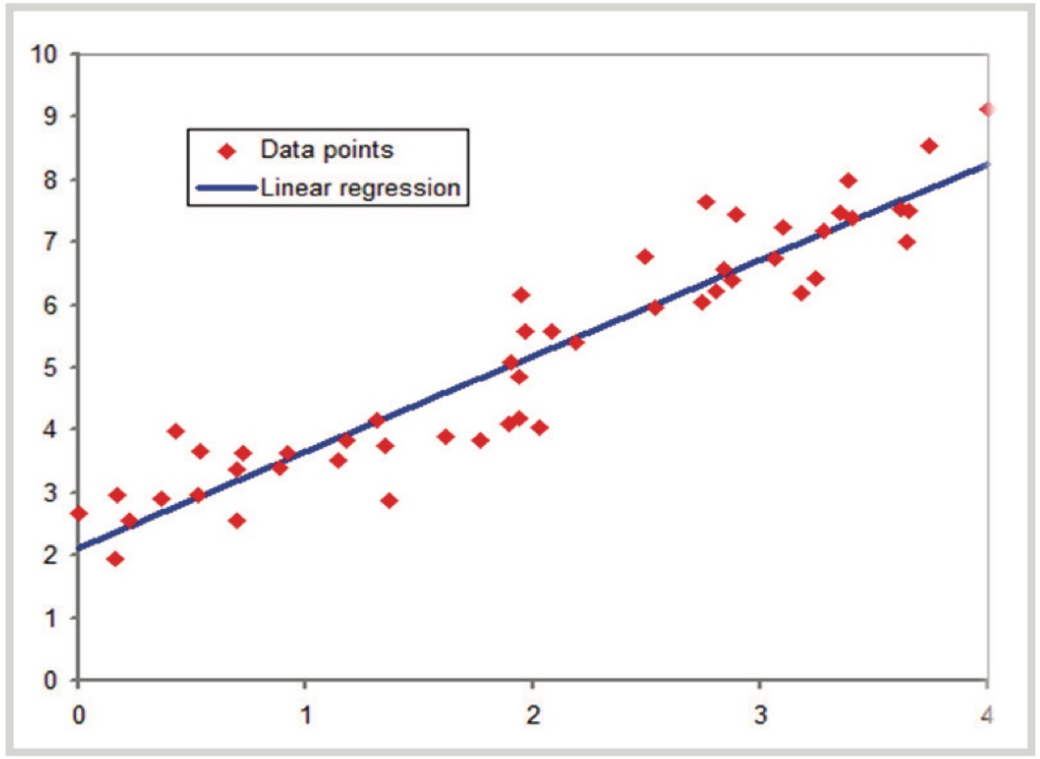
@yeseuly.park – 24.05.19

시험-주요내용 https://www.dataq.or.kr/www/sub/a_07.do#none

빅데이터 모델링	분석모형 설계	분석 절차 수립	분석모형 선정
			분석모형 정의
			분석모형 구축 절차
		분석 환경 구축	분석 도구 선정
			데이터 분할
		분석기법	회귀분석
			로지스틱 회귀분석
			의사결정나무
			인공신경망
			서포트벡터머신
			연관성분석
			군집분석

회귀분석

일반적으로 예측을 목표하는 통계 분석이다. 예측을 하는 방법에 핵심이 되는 개념이 바로 '추세선'이다. 좌표상에서 데이터의 분포와 앞으로의 변화를 가장 잘 설명할 수 있는 하나의 선을 그려내는 것이 회귀분석의 궁극적인 목적이 된다.



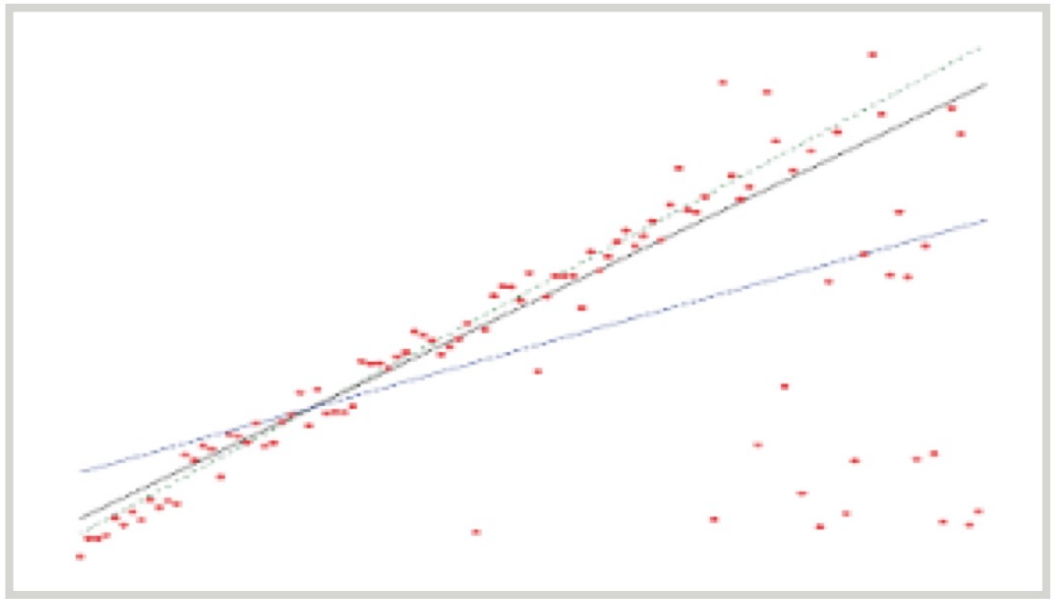
〈직선의 추세선〉

데이터 산포를 통해 가장 예측력이 뛰어난 선을 그린다.

출처 : Wiki

추세선의 의미

직선의 추세선을 수식으로 표현하면, y=ax+b와 같은 1차 방정식이 된다. 여기서 x와 y는 이미 알고 있는 데이터값이다. 여기서 수식을 활용해 데이터 변화의 추세를 확인하는 방법은 x와 y에 들어오는 값을 기준으로 해당 수식을 충족시키는 a(=기울기), b(=절편)값을 찾는 것이다. 즉 추세선을 통한 회귀적 예측이란 곧 a, b를 구하는 과정을 말하는 것이다. 여기서 a와 b를 '회귀계수'라 한다.



〈이상치와 추세선〉

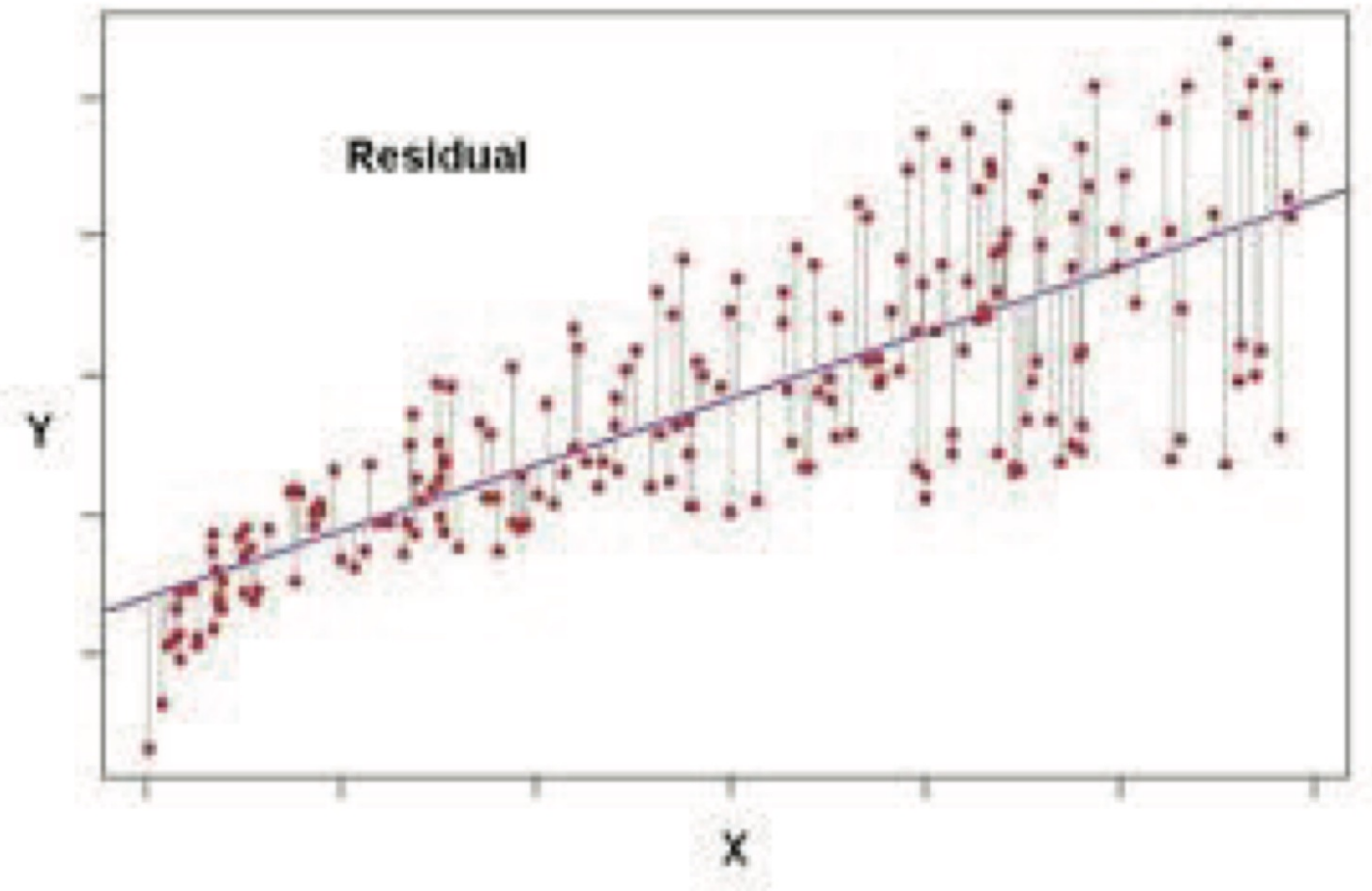
이상치를 제외하면 최적의 회귀계수를 찾아
더욱 높은 예측력을 가지는 회귀선을 그릴 수 있다.

회귀의 시작, 최소제곱법(=최소자승법 Ordinary Least Squares)

그렇다면 최적의 회귀계수 a와 b를 어떻게 찾을 수 있을까. 이때 사용되는 개념이 최소제곱법이다. 최초엔 임의의 추세선(=가설초기화)을 기준으로 분포한 x, y의 좌푯값의 차이(=잔차)를 제곱하여 모두 더한 값이 최소가 되는 지점들을 연결하는 방법이다.

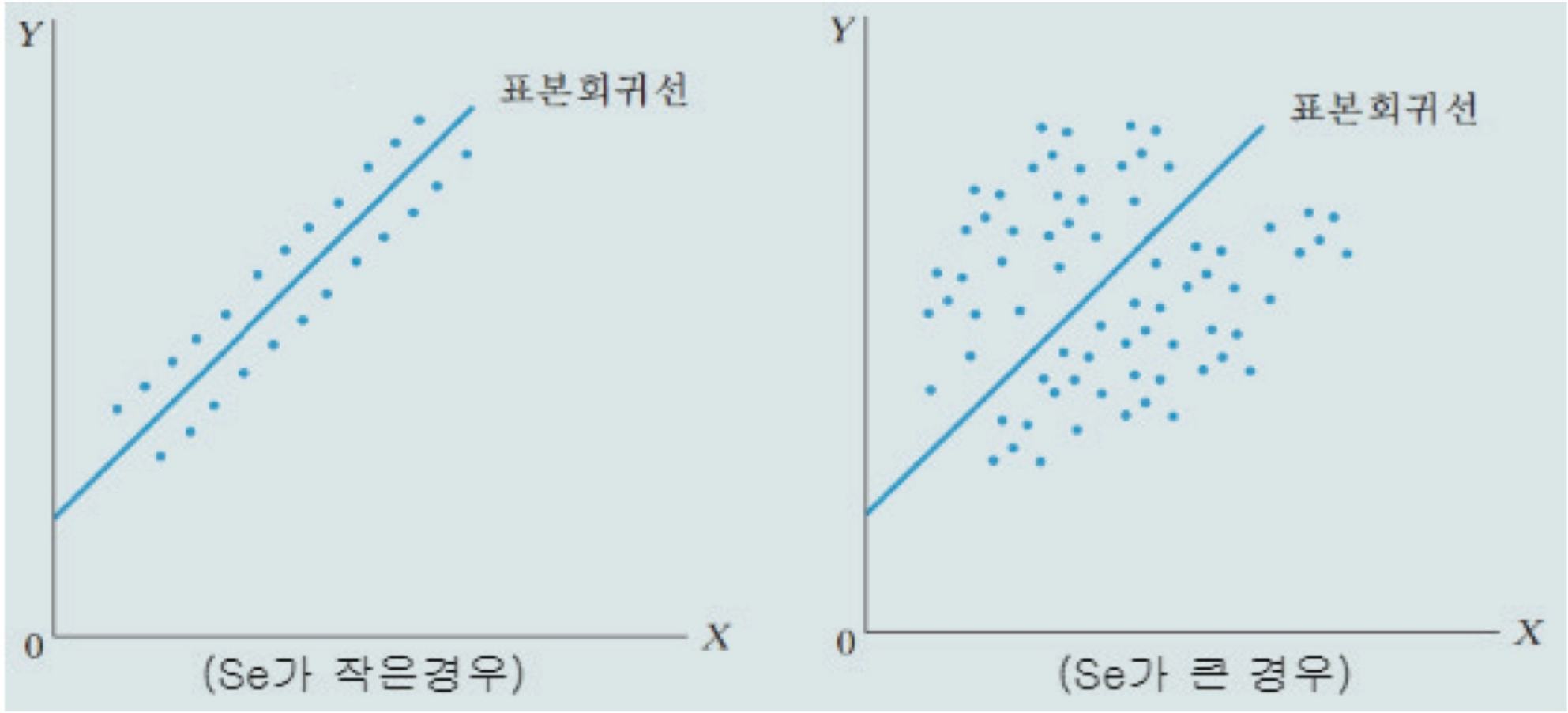
잔차가 최소인 부분을 찾는 이유는 잔차가 곧 '오차'를 의미하기 때문이다. 추세선을 기준으로 실제 분포한 좌푯값과의 거리가 크다면 오차가 큰 것이다. 오차는 추세선보다 클 수(+)도 있고 작을 수(−)도 있다. 값의 크기와 관계없이 차이에 대한 절댓값(=거리)을 필요로 하므로 오차마다 제곱을 하게 되고 그 값들의 합이 최소가 될 때 비로소 추세선과 데이터값과의 오차의 합이 가장 작아지게 된다.

따라서 OLS는 잔차들의 제곱합 값이 가장 작은 값의 선분을 찾아감으로써 평균으로 '회귀'하는 성질을 가지며, 우리는 이를 활용해 회귀적인 분석(예측)을 할 수 있다.



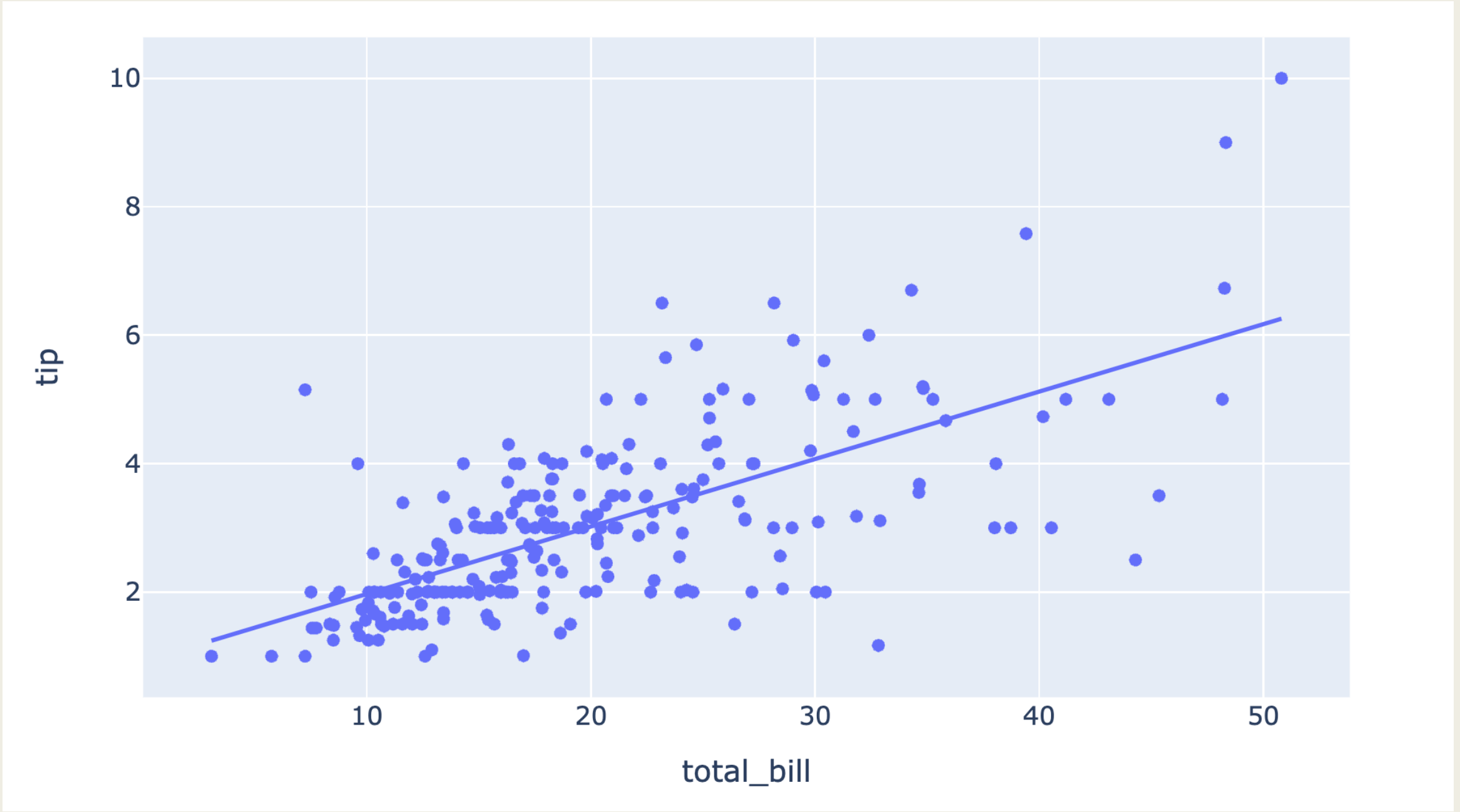
〈잔차와 추세선〉

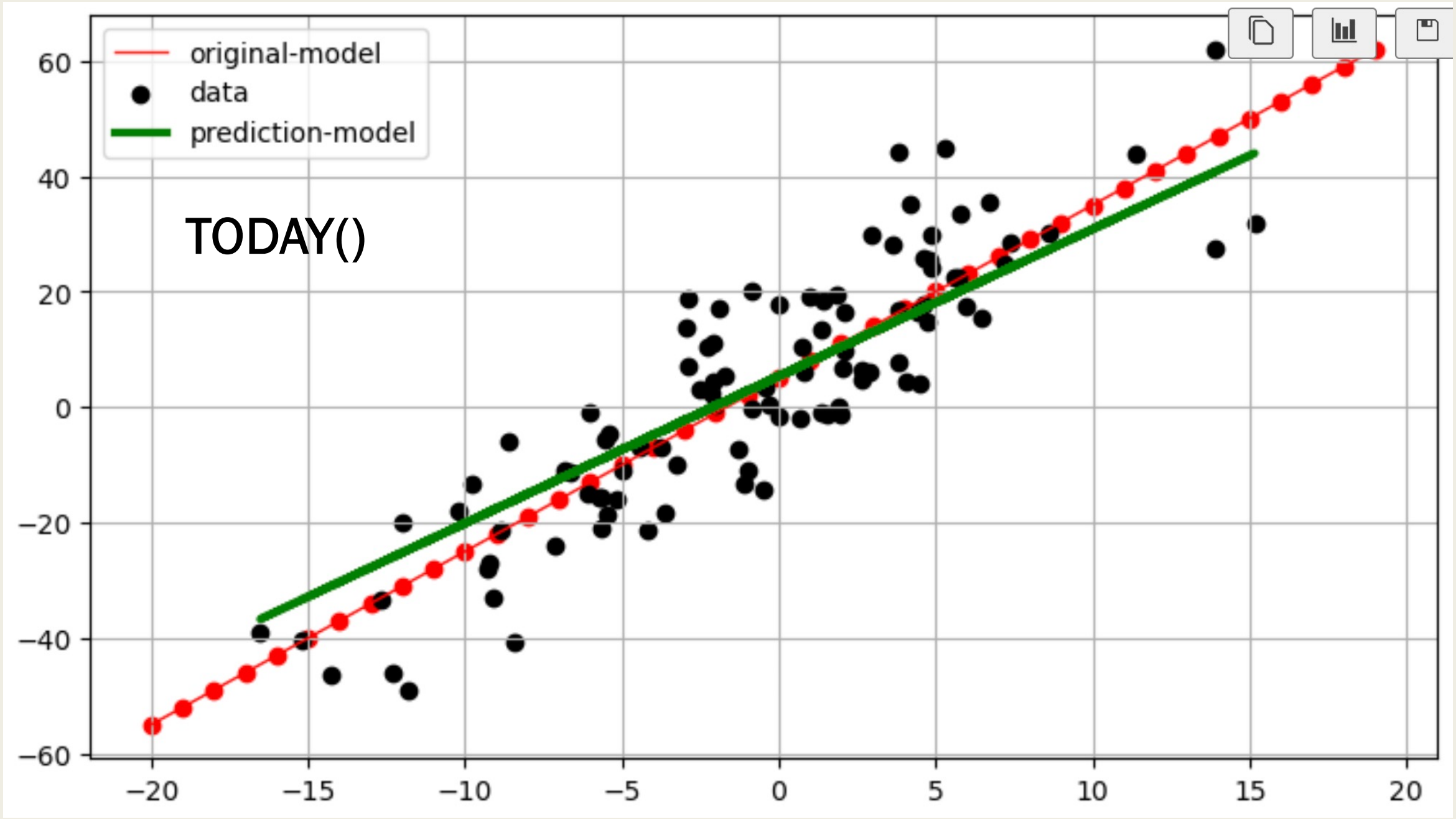
잔차란 곧 데이터값과 추세선과의 오차이고 각 오차들의 제곱합이 최소가 될 때 비로소 최적의 추세선을 찾을 수 있다.



〈표준오차와 추세선〉

표준오차(SE)가 크면 y값을 추정하는 회귀계수(=추세선)의 우연정도(P-value)가 커지므로 해당 추세선(=회귀식)의 설명력은 떨어질 수밖에 없다.


10
8
6
4
2
tip
10
20
30
40
50
total_bill


original-model
data
prediction-model
TODAY()
60
40
20
0
−20
−40
−60
−20
−15
−10
−5
0
5
10
15
20