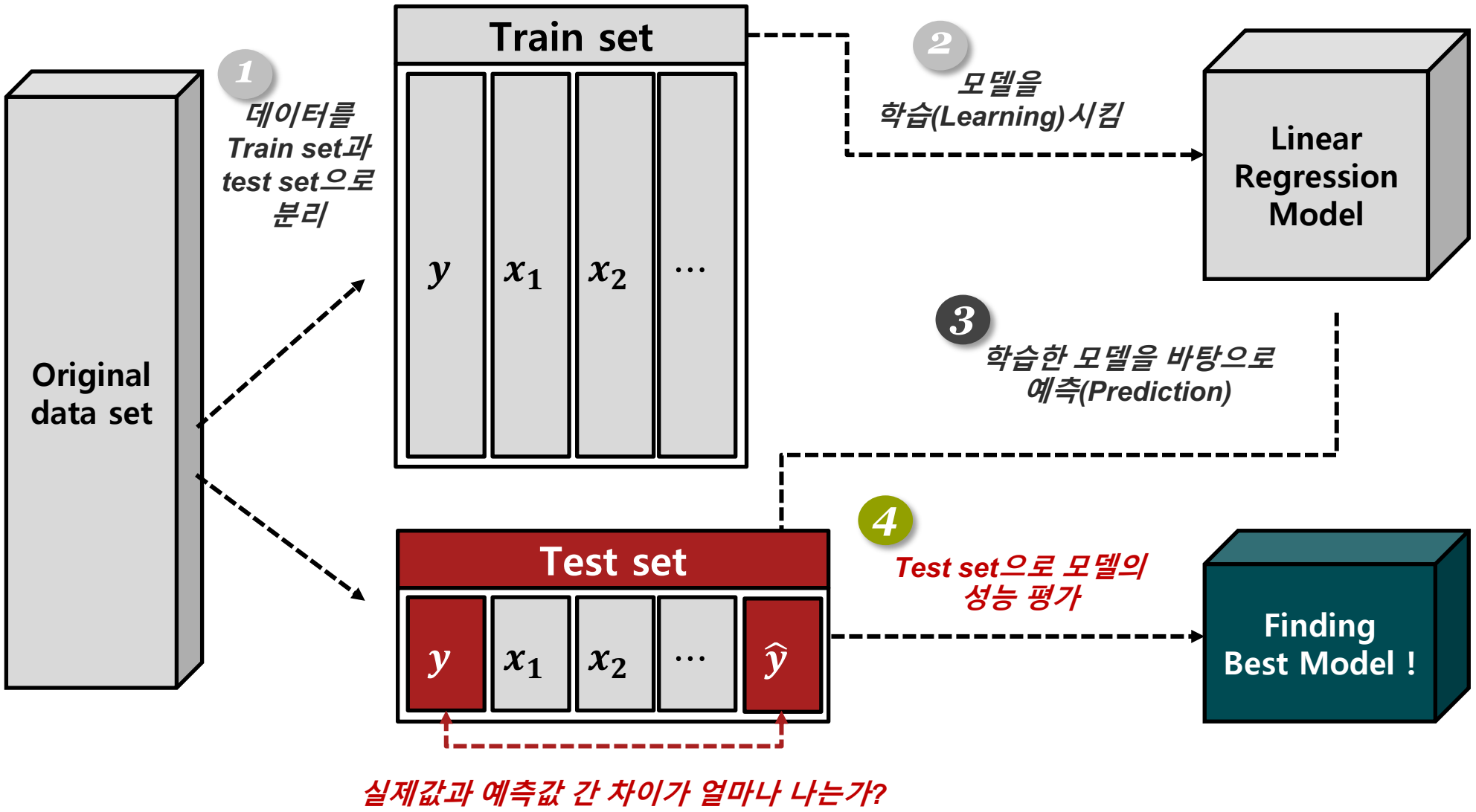
Ⅱ. 일반 회귀분석 (Regression)

❷ Linear RegressionⅡ

선형회귀분석 Review

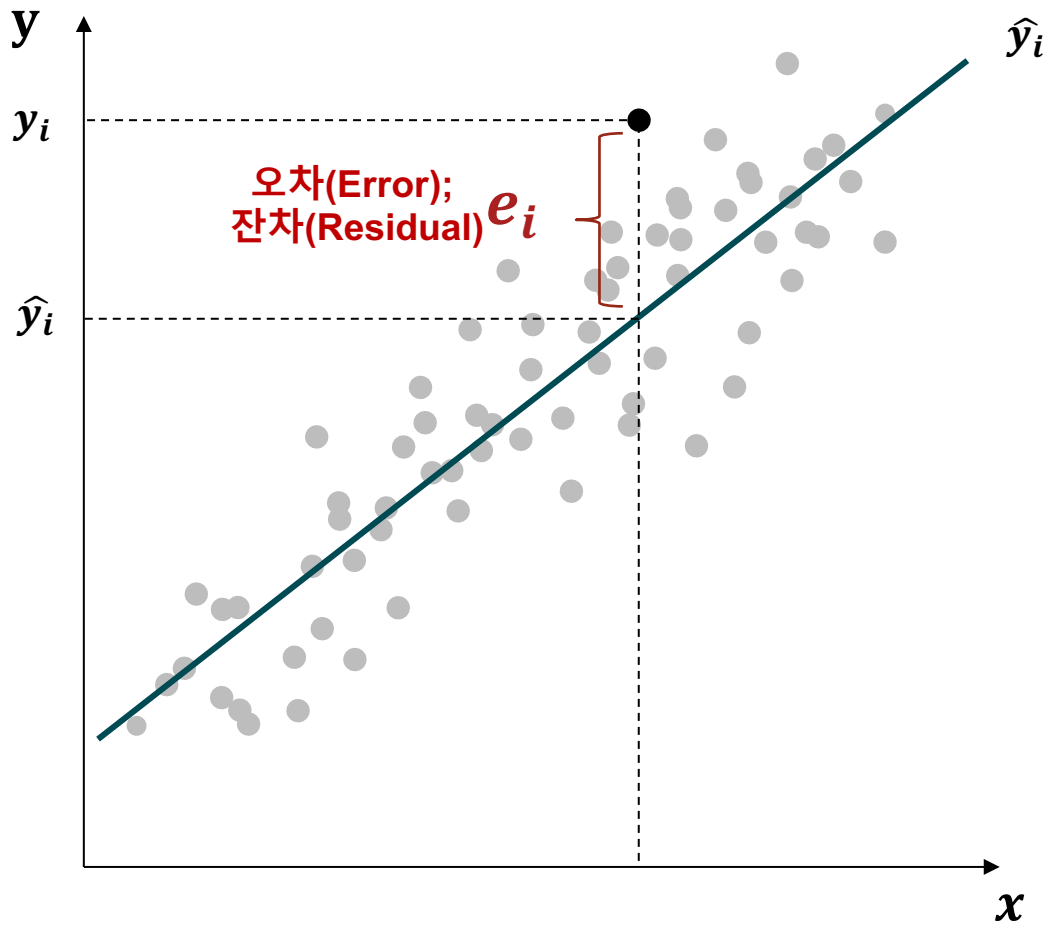


선형회귀모형의 기본가정

- ✓ **잔차 (또는 오차항)는 모든 독립변수 값에 대해 동일한 분산을 갖는다.**
- ✓ **잔차 (또는 오차항)의 평균(기대값)은 0이다(실제 데이터로 잔차를 계산하면 0에 근접하게 나온다).**
- ✓ **수집된 자료의 잔차 (또는 오차항) 는 정규분포를 이루고 있다.**
- ✓ **독립변수 상호간에는 상관관계가 없어야 한다(완전히 없는 경우는 드물지만 비교적 낮아야 한다).**

왜 실제값과 예측값 간 차이가 발생하는 것인가?

선형 회귀모형은 가장 잘 설명할 수 있는 하나의 직선을 긋는 것이므로 각각의 점으로부터 거리가 발생하게 되고, 이 거리가 오차(Error)가 되어 실제값과 예측값 사이에 차이가 발생함



$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Best model = The smallest error

각각의 점에서 발생한 오차(Error)들을 평균내면 모형의 평균오차를 구할 수 있고, 여러 대안들이 있다면 평균오차가 가장 적은 모형이 **가장 좋은 모형**이 될 수 있음

오차(Error)의 종류

회귀분석의 잔차(Residual)의 합은 항상 0 이므로 단순 평균을 내면 0이 나옴. 따라서, 모형평가 지표로는 평균절대오차(MAE) 또는 제곱근평균제곱오차(RMSE)가 활용됨

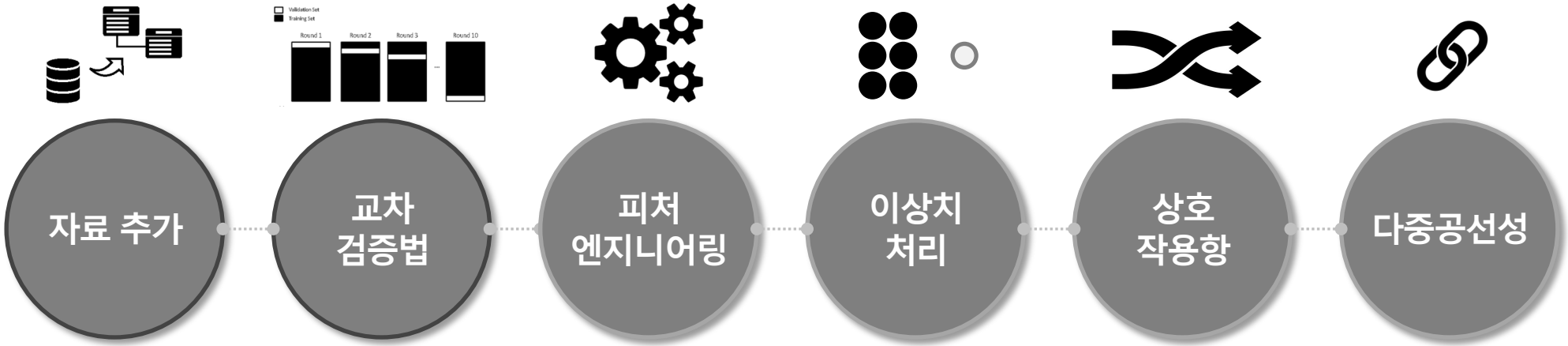
평균절대오차(Mean Absolute Error, MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\, y_i - \hat{y}_i \,| = \frac{1}{n}\sum_{i=1}^{n}|\, e_i \,|$$

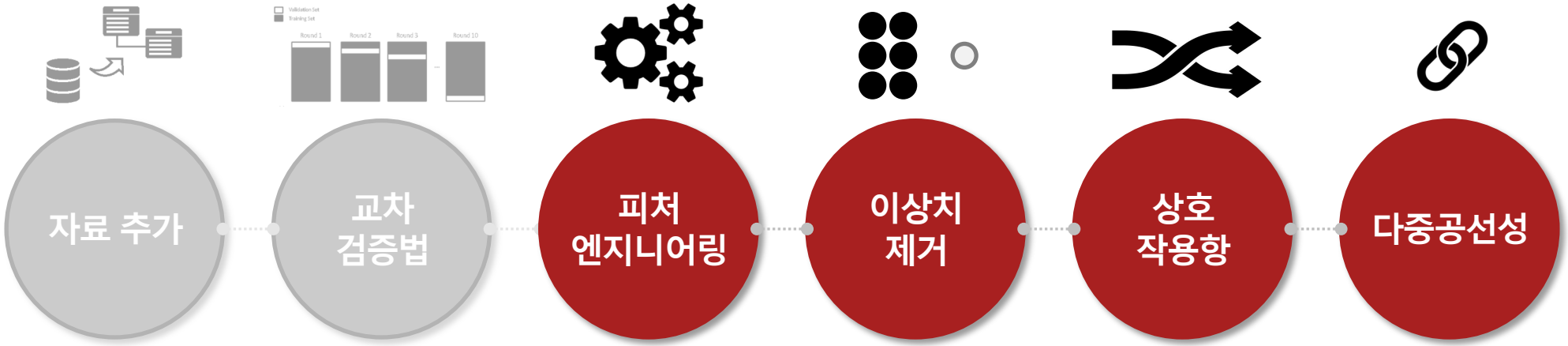
제곱근평균제곱오차(Root Mean square Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^{\,2}}$$

모형개선(Model Improvement)



모형개선(Model Improvement)



"추가적인 시간 및 비용이 소요될 수 있음"

모형개선#1 – 피처엔지니어링 (Feature Engineering)

피처 엔지니어링은 주어진 피처(Feature)들을 이용해 해당 도메인에 대한 지식 및 특성 등을 미리 알거나 탐색적 분석을 통해 알게 된 사실을 바탕으로 **유의미한 변수를 생성, 선택 및 변환하는 과정**

거래처	A 제품 매출액	B 제품 매출액	C 제품 매출액	…
이마트	100,000	60,000	230,000	…
롯데마트	200,000	-30,000	480,000	…
홈플러스	150,000	40,000	-110,000	…
⋮	⋮	⋮	⋮	⋮

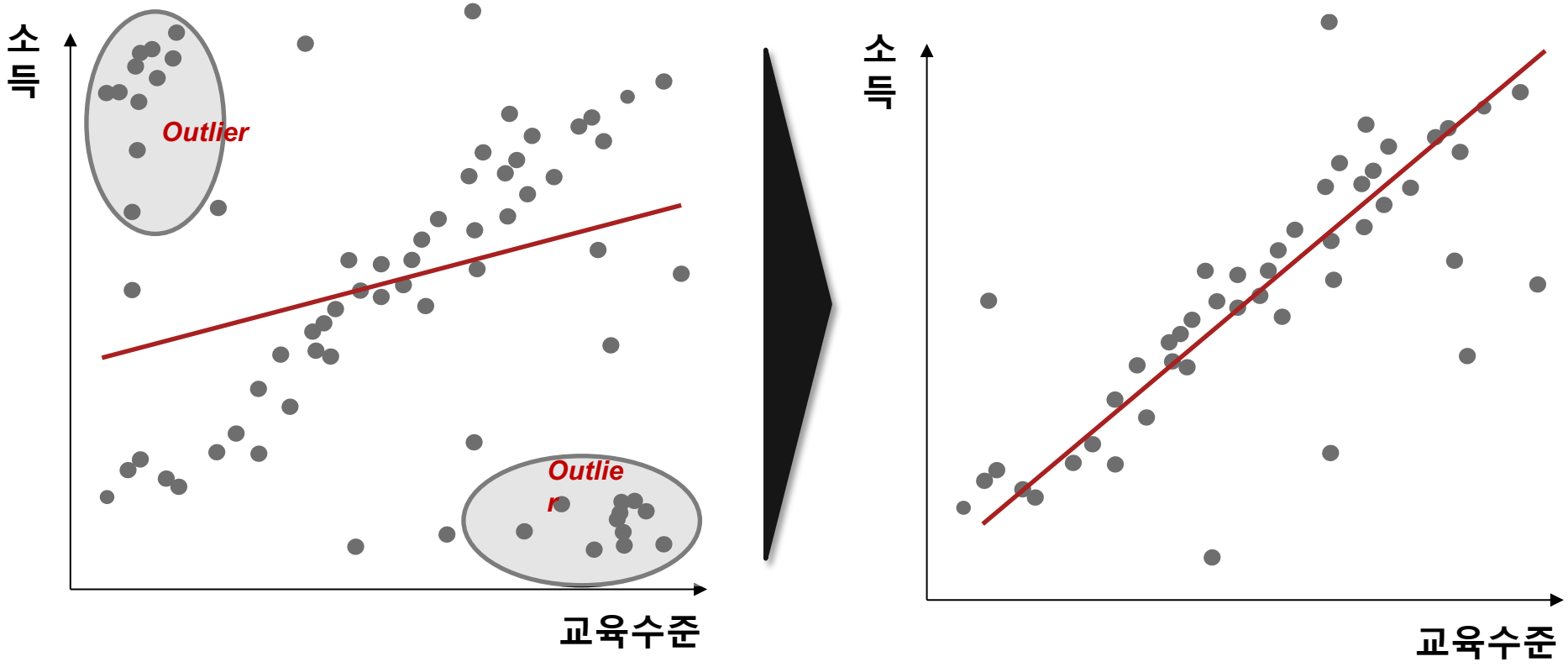
환입으로 인해 변수의 값이 음수(-)인 값이 존재함

거래처	A 제품 매출액	B 제품 매출액	C 제품 매출액	…
이마트	100,000	60,000	230,000	…
롯데마트	200,000	0	480,000	…
홈플러스	150,000	40,000	0	…
⋮	⋮	⋮	⋮	⋮

사전 지식을 바탕으로, 매출액이 음수(-)인 값은 모두 0으로 바꿈

모형개선#2 – 이상치 발견 (Outlier detection)

이상치(Outlier)는 데이터 상에 존재는 하지만 일반적이지 않은 분포나 형태를 나타내고 있는 자료를 의미하며, 자료의 수가 많으면 이상치의 영향이 줄어드나 이상치가 과도하게 많으면 모형성능이 떨어짐



모형개선#2 – 이상치 발견 (Outlier detection)

이상치 (Outlier)

제거

이상치를 제거하면 모형 성능을 개선할 수도 있지만 이상치가 존재하는 관측치 자체가 모형에 투입되지 못하므로 정보의 손실 발생

모형에 투입되지 못함

ID	소득	나이	…
1	350	42	…
2	400	*339 (outlier)*	…
⋮	⋮	⋮	⋮

조정

이상치를 조정하는 방법은 여러가지가 있으며, 대표적으로 해당 변수의 분포 중 양측 90%에 해당하는 데이터로 이상치를 끌어오는 방법이 있음

모형에 투입가능

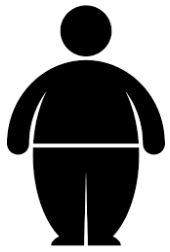
ID	소득	나이	…
1	350	42	…
2	400	*45*	…
⋮	⋮	⋮	⋮

모형개선#3 – 상호작용항 (Interaction Term)

상호작용효과는 서로 독립적인 변수이지만 독립적인 두 변수가 각각 미치는 영향 외에도 두 변수가 동시에 발생했을 때, 시너지(Synergy) 효과가 발생하 는 경우를 말하며 모형에 이를 상호작용항으로 반영함

흡연자

“흡연자이면서 비만 환자라면? ”



비만

보험료(Insurance expense)

모형개선#4 – 다중공선성 (Multicollinearity) 제거

다중공선성은 설명변수로 들어가는 두 변수 간 상관관계가 매우 높은 경우를 말하며, 이 경우 다중공선성을 제거해야 독립변수의 진정한 영향정도를 파악할 수 있고 모형 성능을 높일 수 있음