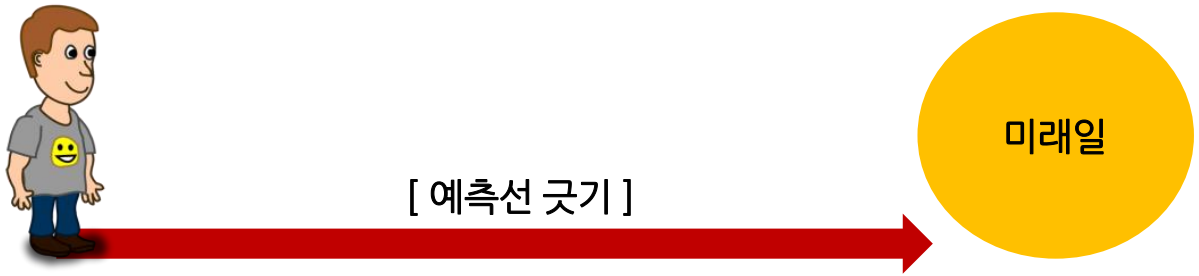
Regression (회귀)

회귀분석의 종류 (1) : 선형회귀

- 가장 훌륭한 선 긋기 → 머신러닝은 미래의 방향을 설정하는 것에서 부터 시작합니다.
- 선의 방향을 잘 정하면, 그 선을 따라가는 것만으로도 지금은 보이지 않는 미래의 것을 예측 가능합니다.



회귀분석의 종류 (2) : 로지스틱 회귀

- 전달받은 정보를 놓고 참과 거짓 중에 하나를 Output으로 선택하는 방법론입니다.
- 참, 거짓 판단장치 라고도 합니다 → 이진 분류에 많이 사용합니다.
- 예제) 합격자 발표에서, 점수와 상관없이, '합격'과 '불합격'만 존재합니다.

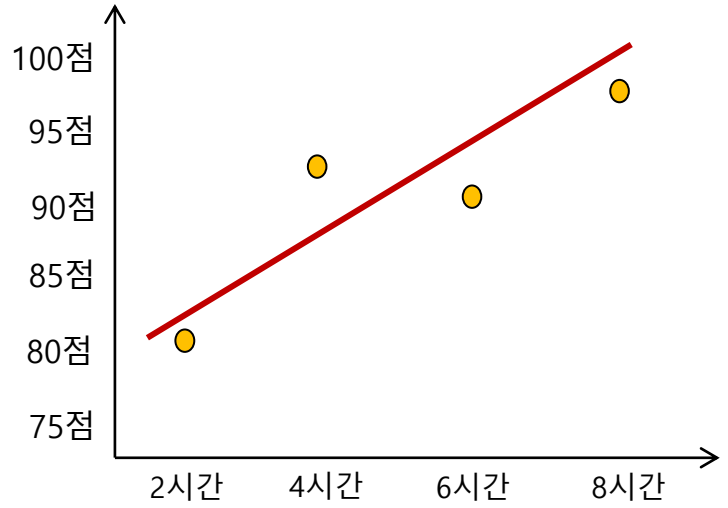
공부한 시간	2	4	6	8	10	12	14
합격여부	불합격	불합격	불합격	합격	합격	합격	합격

선형회귀: 가장 훌륭한 선 긋기 (1)

- Y=aX +b로 표현 될 수 있습니다.
  - ‘X값이 변함에 따라 Y값도 변한다.’ → Simple Linear Regression

예) 독립변수, ‘X’가 공부한 시간, 성적 ‘Y’를 예측할 경우, ‘X’ 가 한 개 이므로, ‘Simple Linear Regression’.

공부한 시간	2 시간	4 시간	6 시간	8 시간
성적	81점	93점	91점	97점



선형회귀: 가장 훌륭한 선 긋기 (2)

- 우리의 목표: 가장 정확한 선 긋기!!
  - 가장 정확한 기울기 a와 절편 b를 찾으면 된다.
  - 여러 가지 선을 그을 수 있고, 여러 가지 선 중, 반복되는 선 긋기를 통해서, 가장 훌륭한 선을 찾는다.
  - 선형회귀는 임의의 직선을 그어 이에 대한 평균 제곱 오차를 구하고, 이 값을 가장 작게 만들어 주는 a와 b를 찾아가는 작업이다.

- [How] 어떻게 가장 훌륭한 선 찾을까? 오차(예측값 – 실제값) 줄이기
  - 가장 많이 쓰는 방법: 평균제곱오차(MSE: Mean Square Error)
    - $MSE = 1/n \sum(\text{예측값} - \text{실제값})^2$

선형회귀 : 가장 훌륭한 선 찾는 기준: 예측 모델 성능 평가

1. 가장 많이 쓰는 방법: 평균 제곱 오차(MSE: Mean Square Error)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

2. 평균 제곱근 오차(RMSE: Root Mean Square Error) : MSE 값은 오류의 제곱을 구하므로, 실제 오류의 평균보다 값이 더 커지는 특성이 있을 수 있으므로, MSE에 루트를 씌운 경우를 의미합니다.

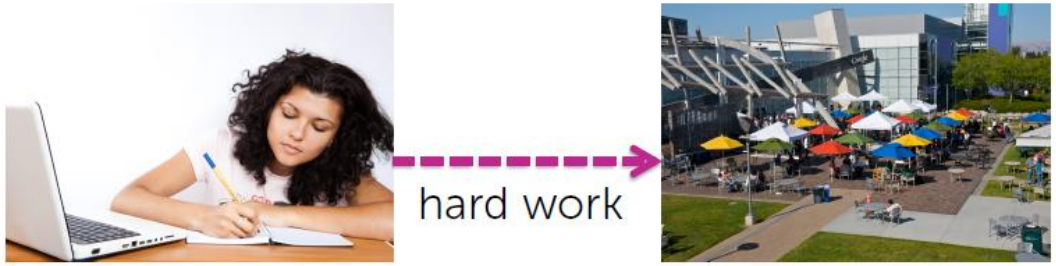
$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

3. 평균절대 오차 (Mean Absolute Error: 평균절대 오차) : 실제 값과 예측 값의 차이를 절대값으로 변환해 평균한 것

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

선형회귀: Multiple Linear Regression 예제

- 'X' 값 한 개 만으로 정확히 설명할 수 없을 경우는, X 를 여러 개 놓을 수 있습니다. → X_1, X_2 ,X_3 …(Multiple Linear Regression)



$$\hat{y} = \hat{w}_0 + \hat{w}_1 \text{ performance} + \hat{w}_2 \text{ capstone} + \hat{w}_3 \text{ forum}$$

informed by other students who completed specialization

[데이터사이언스를 공부한 후, 나의 연봉은?]

(X_1) : Recent history of stock price

(X_2) : News events

(X_3) : Related commodities

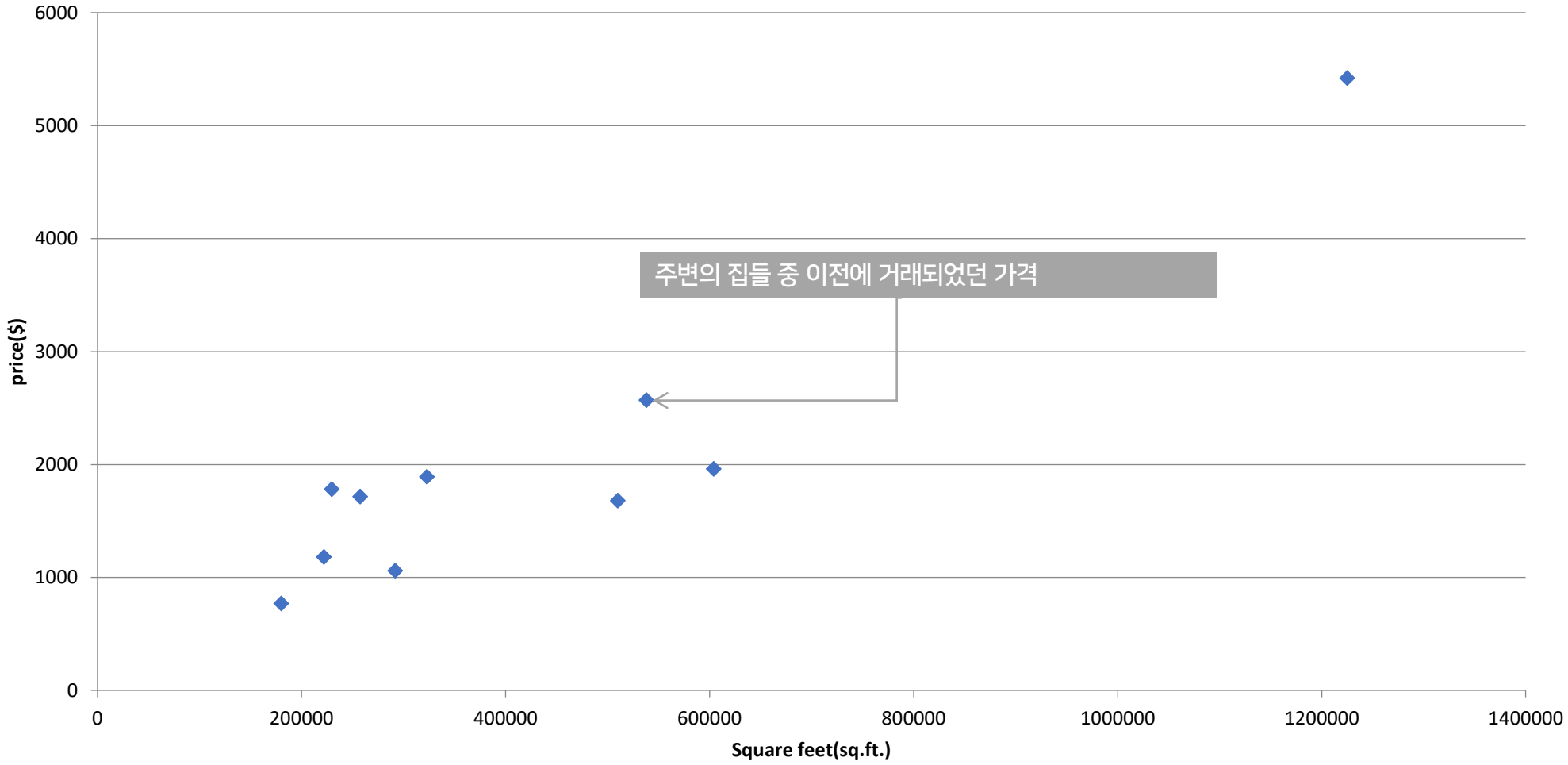
[주식시장 예측]

예제로 배우기: 집값 예측 (1)

최근 주변의 부동산 시세를 살펴본다. 보통 얼마에 거래가 되었을까?

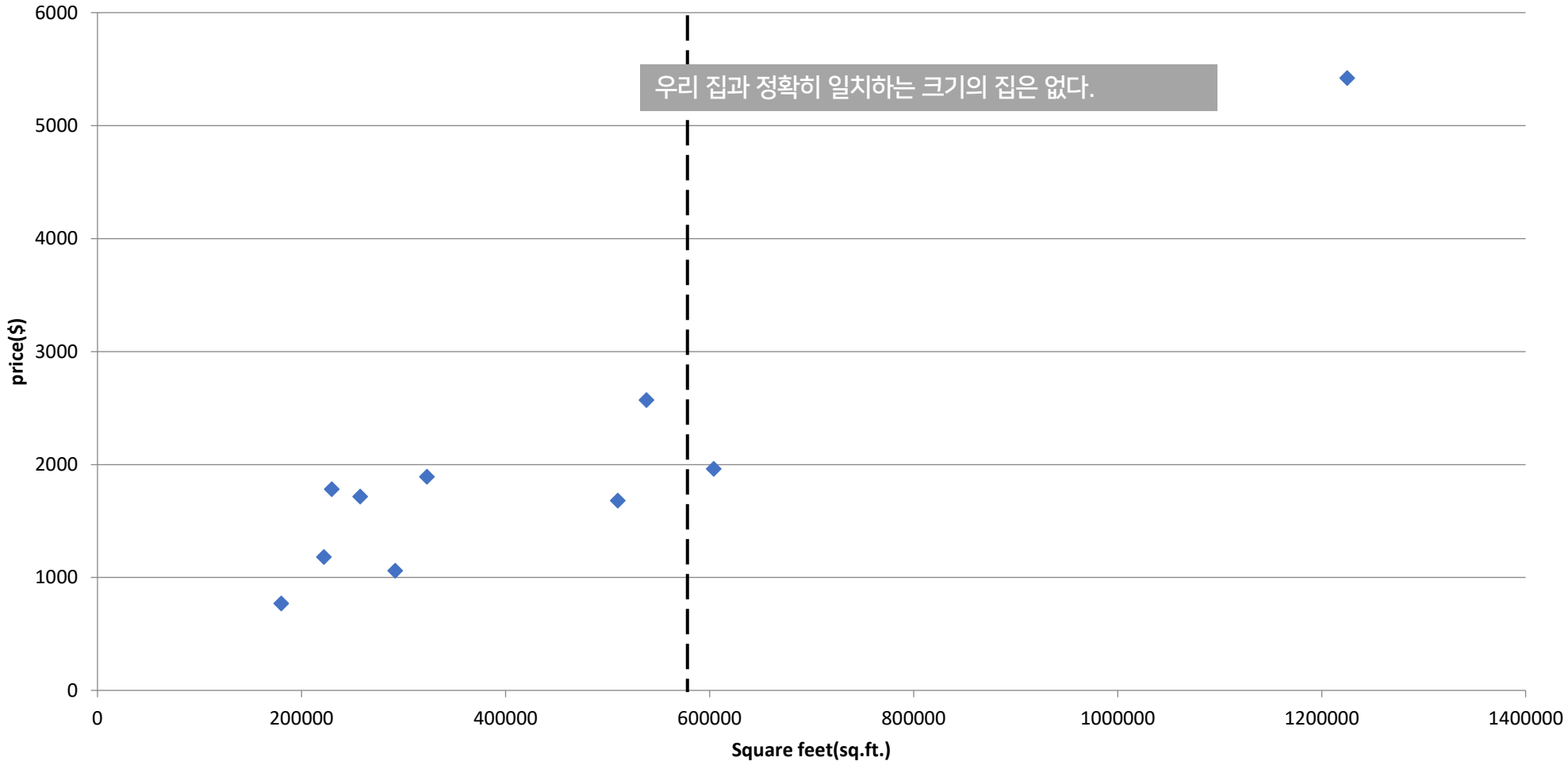
- 방법: 여러가지 특징 세트(Features, Attributes, Columns, Xs)가 있을 때, 특징의 변화에 따라, Output(Y) 의 변화를 살펴본다.
- 주택가격 예측을 통한 사례:
  - 주택크기, 침실 개수, 화장실 개수 → Feature Set (Xs)
  - 관찰을 통해 집값 예측 (Y)

예제로 배우기: 집값 예측 (2)



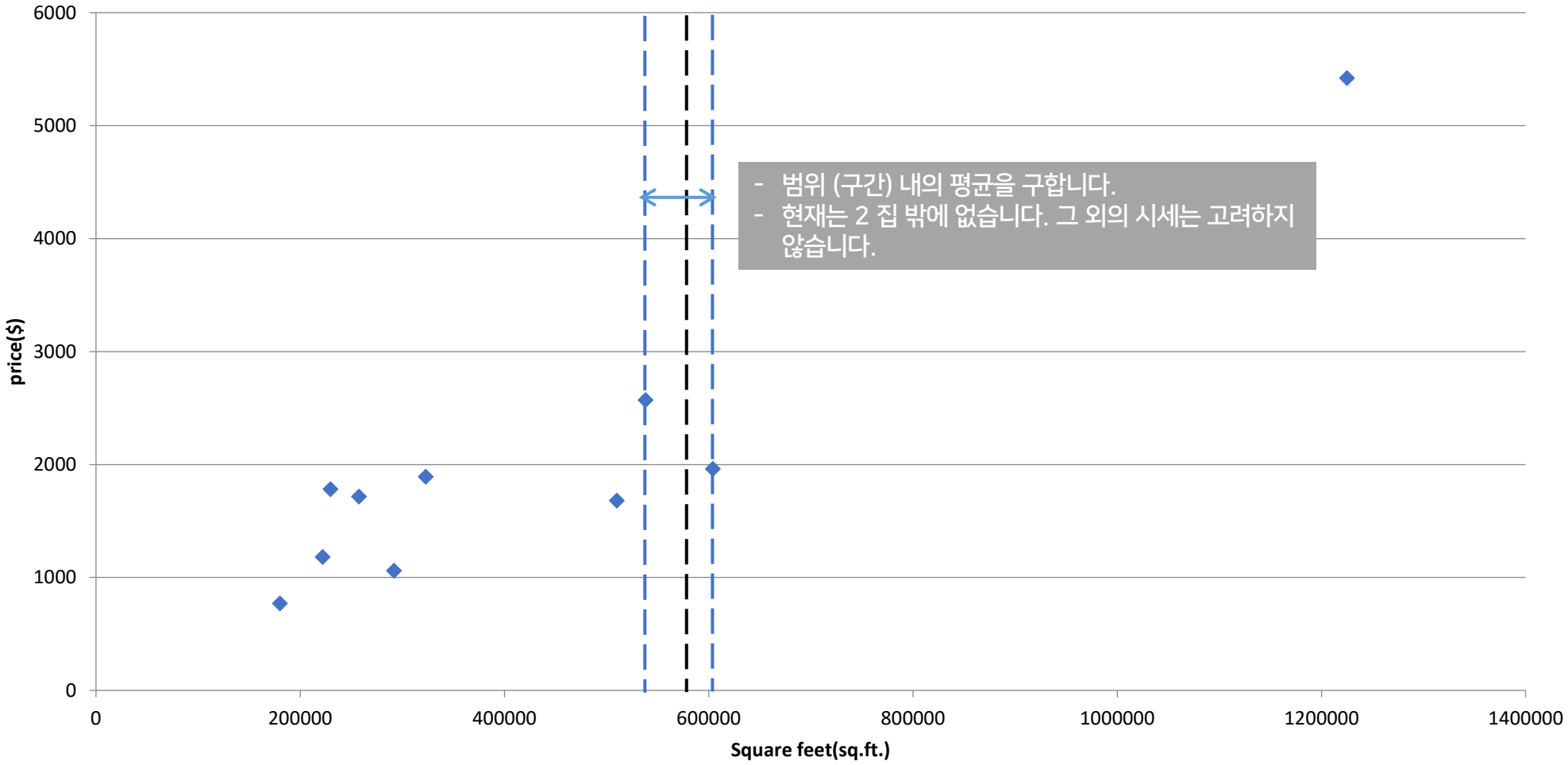
X: Square Feet
Y: Price

예제로 배우기: 집값 예측 (3)



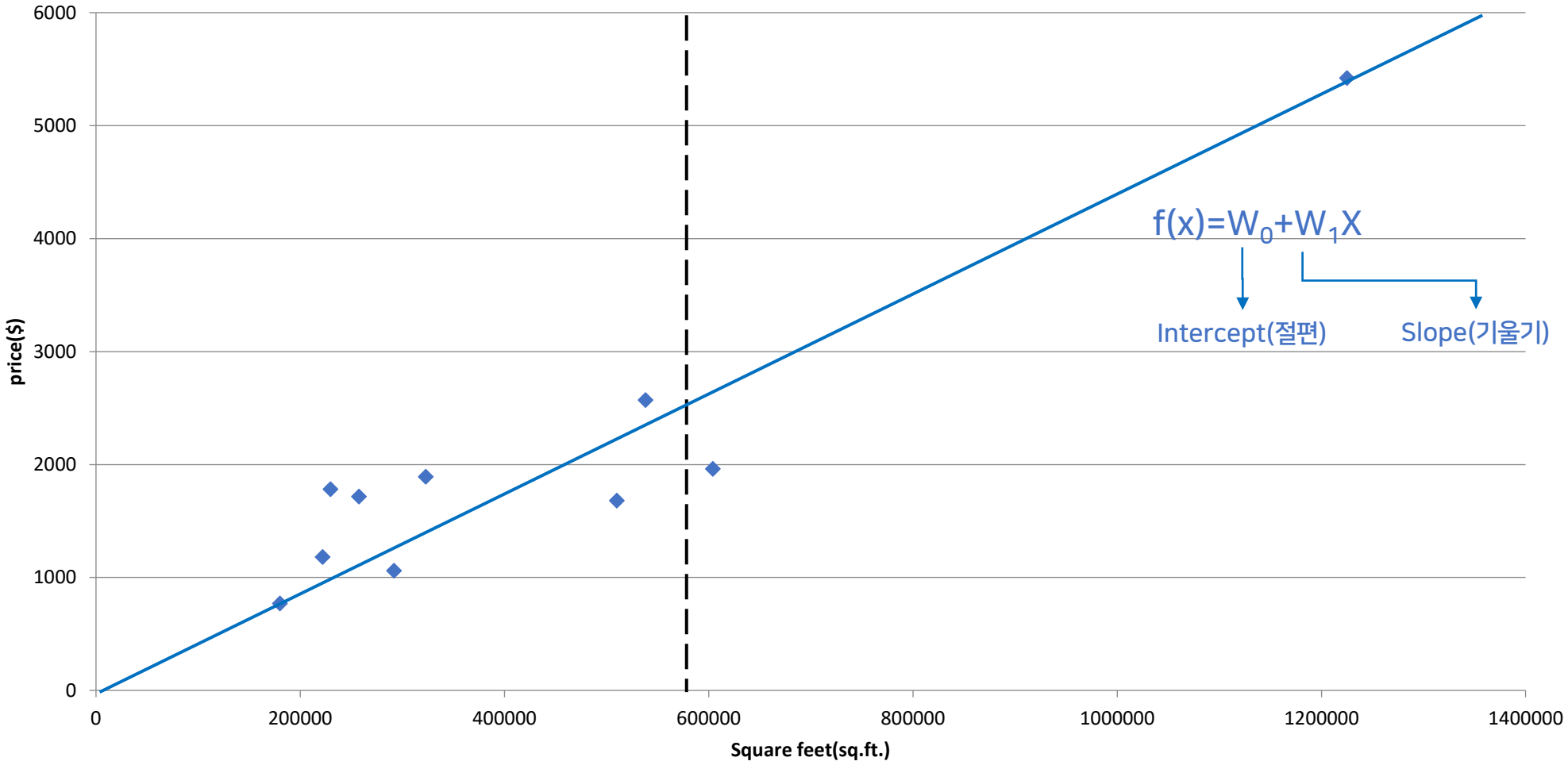
X: Square Feet
Y: Price

예제로 배우기: 집값 예측 (4) 주변 시세를 이용하기



X: Square Feet
Y: Price

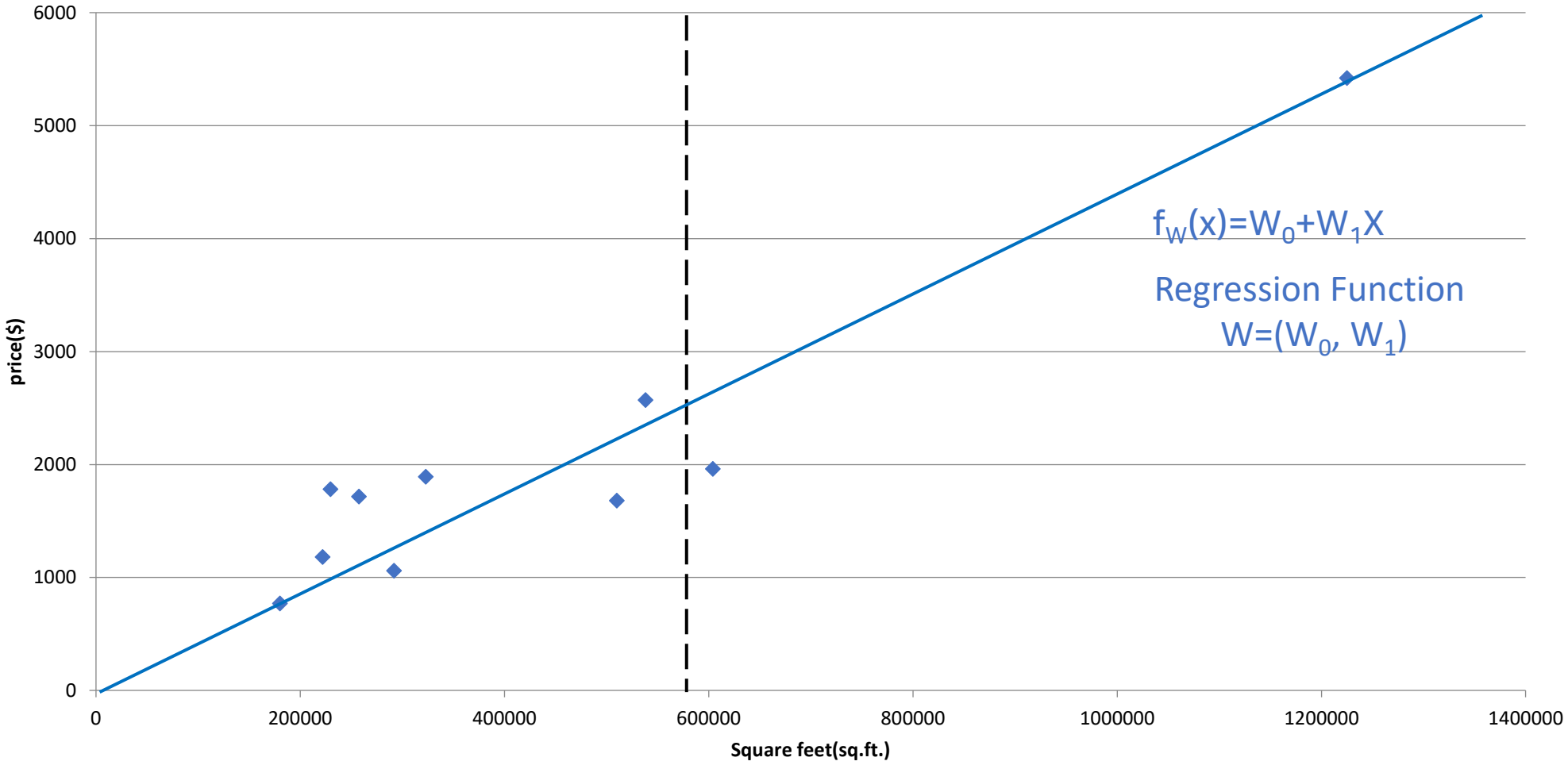
예제로 배우기: 집값 예측 (5) 모든 집값을 다 이용해서 예측하기



X: Square Feet
Y: Price

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

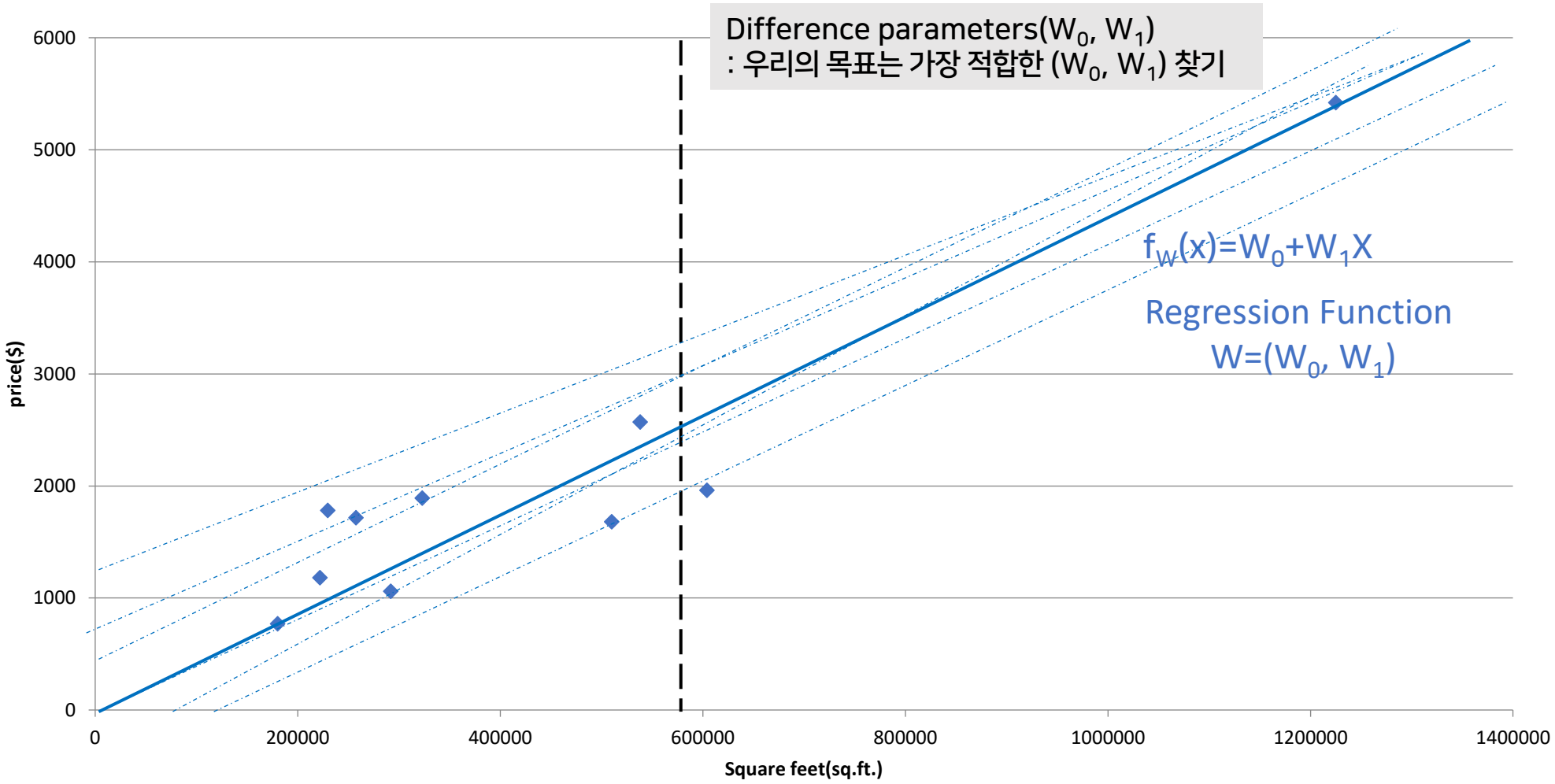
예제로 배우기: 집값 예측 (6) 모든 집값을 다 이용해서 예측하기



X: Square Feet
Y: Price

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

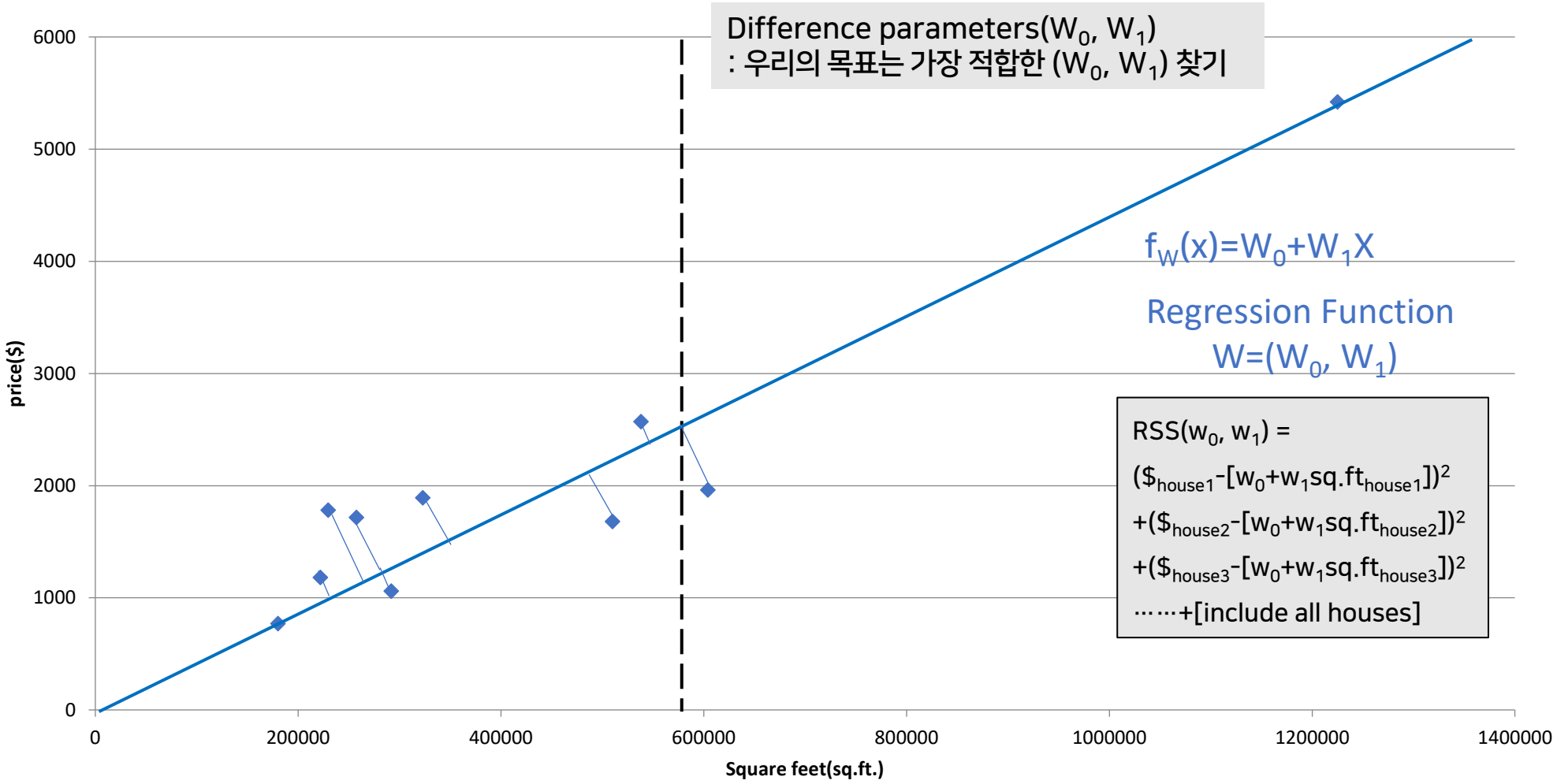
예제로 배우기: 집값 예측 (7) 가장 훌륭한 선 긋기



X: Square Feet
Y: Price

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

예제로 배우기: 집값 예측 (8) RSS (Residual Sum of Squrares 이용하기)



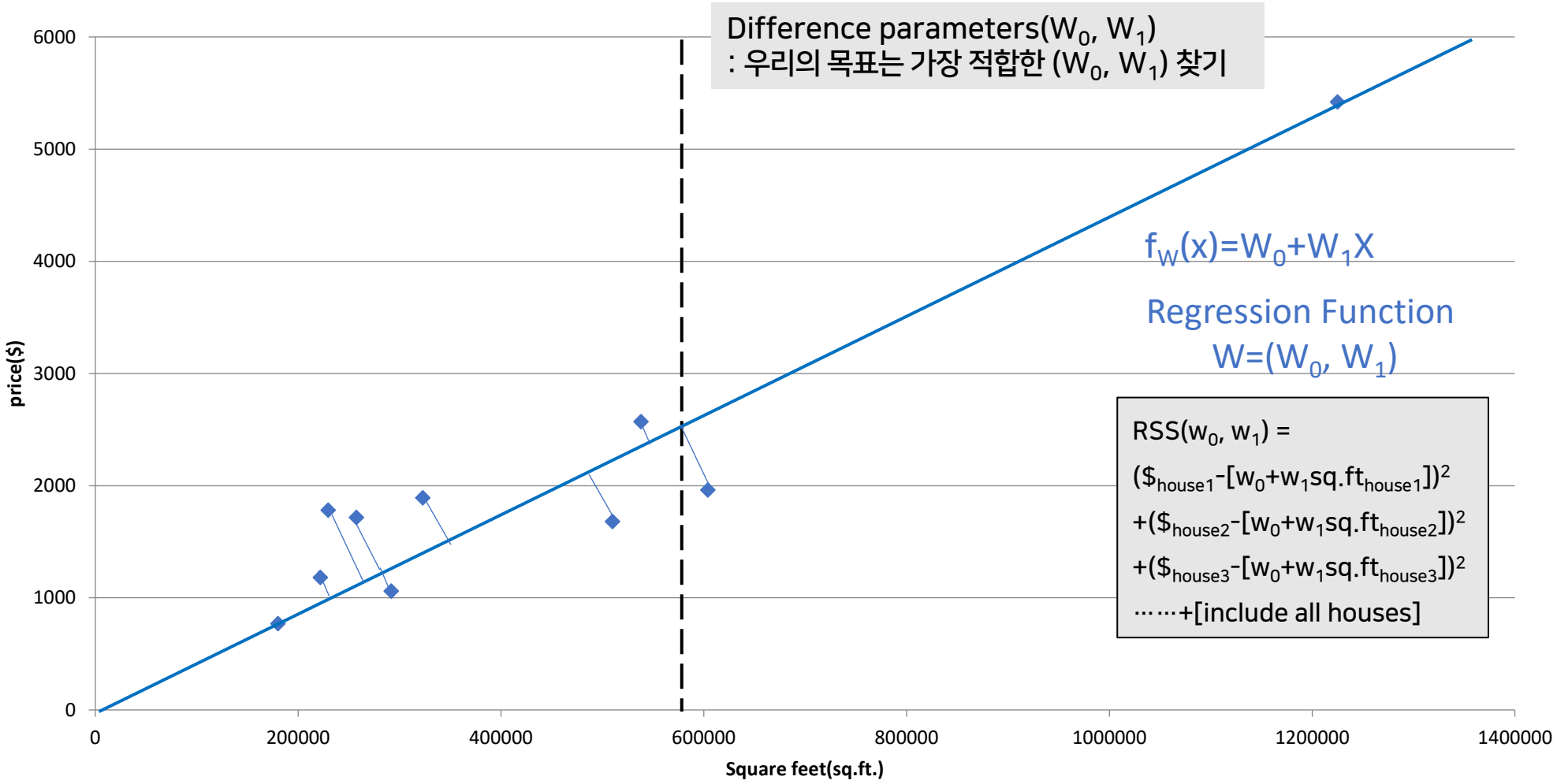
X: Square Feet
Y: Price

RSS = N* MSE

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

예제로 배우기: 집값 예측 (9) RSS(W_0, W_1)이 가장 작은 것 == Best Line == 가장 훌륭한 직선



X: Square Feet
Y: Price

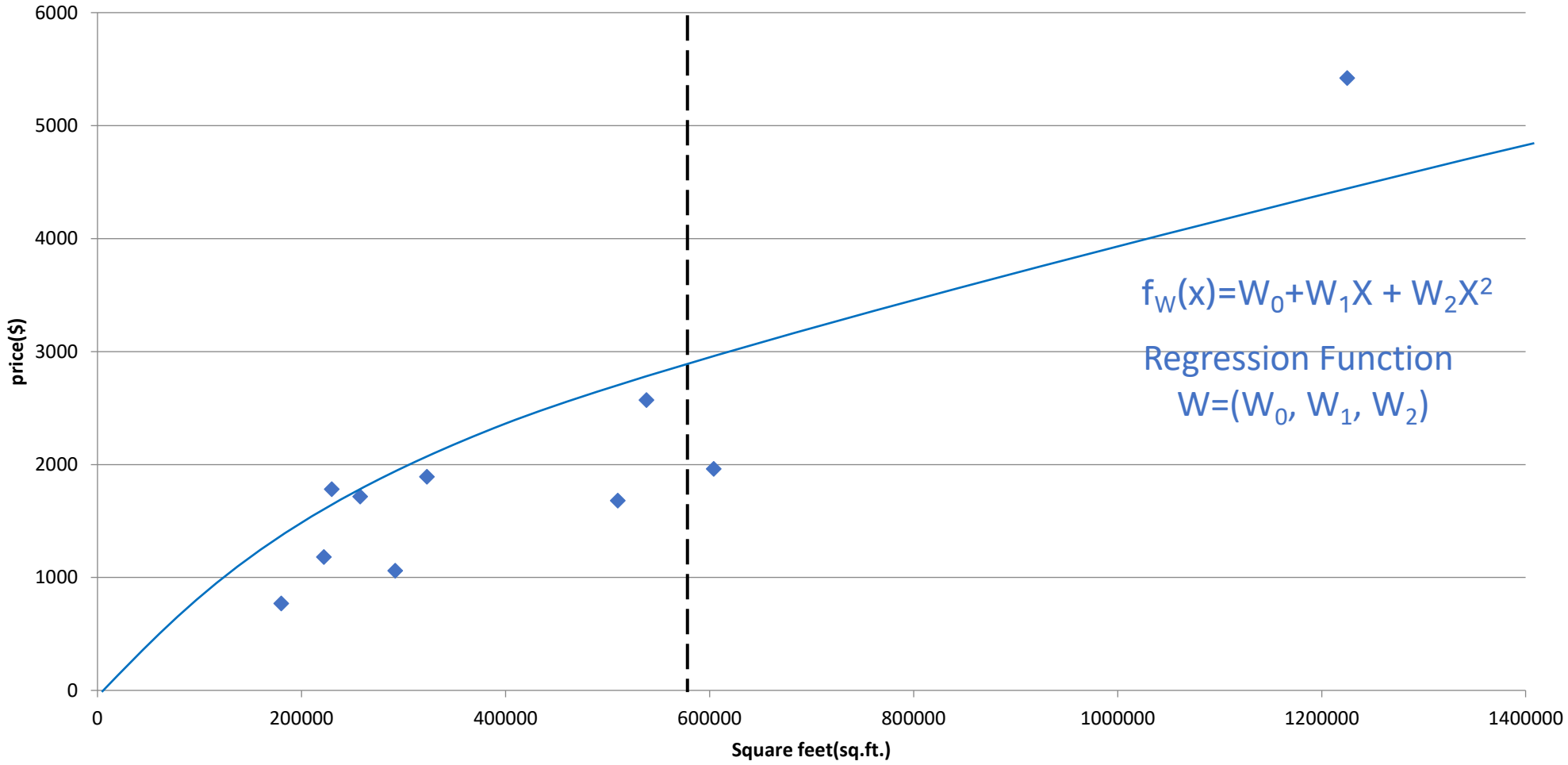
RSS = N* MSE

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

그렇다면,
꼭 직선이어야만 할까요?

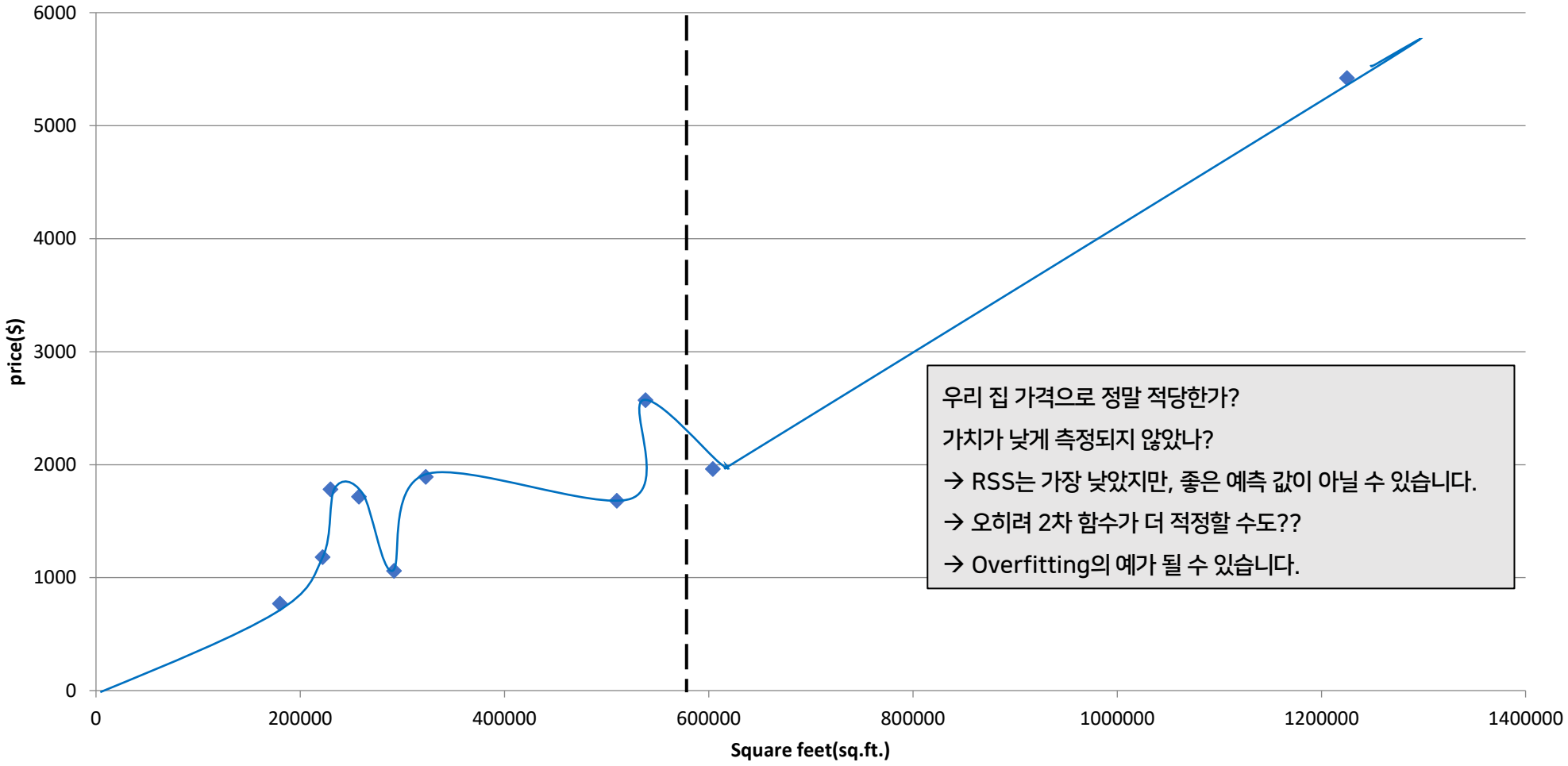
선형은 직선을 의미하는 것이 아니라, 계수의 곱과 합으로 이루어진 것을 의미합니다.



X: Square Feet
Y: Price

2차 함수 곡선으로 구할 수도 있습니다. 가장 훌륭한 선이 곡선 일 수 있습니다.

모든 것을 정확히 고려하면, Overfitting이 될 수 있습니다.



X: Square Feet
Y: Price

2차 함수 곡선으로 구할 수도 있습니다. 가장 훌륭한 선이 곡선 일 수 있습니다.

Regression 을 모델링 하기

- 데이터 세팅
- 학습하기
- 평가하기 with RSS

데이터 셋 : Training Data Set + Test Data Set

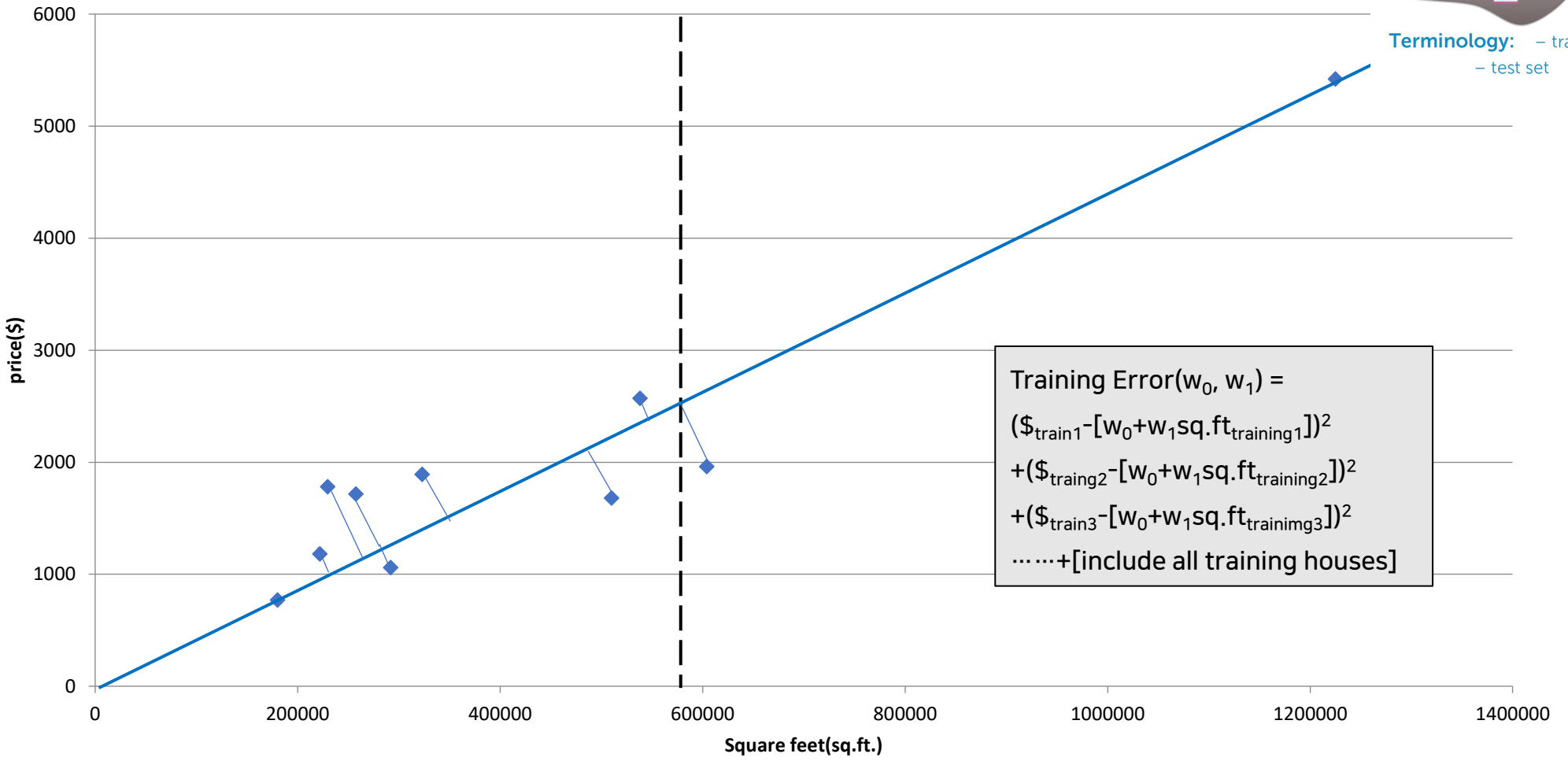
- 모델링을 위해서, Training Set과 Test Set으로 나눕니다.
- Training Set으로 모델을 만든 후, Test Set으로 모델을 검증합니다.

Terminology: – training set

– test set

Training Error

Terminology: – training set
– test set



Training Error(w_0, w_1) =
($\$_{train1}$-[$w_0$+$w_1$sq.ft$_{training1}$])^2
+($\$_{traing2}$-[$w_0$+$w_1$sq.ft$_{training2}$])^2
+($\$_{train3}$-[$w_0$+$w_1$sq.ft$_{trainimg3}$])^2
……+[include all training houses]

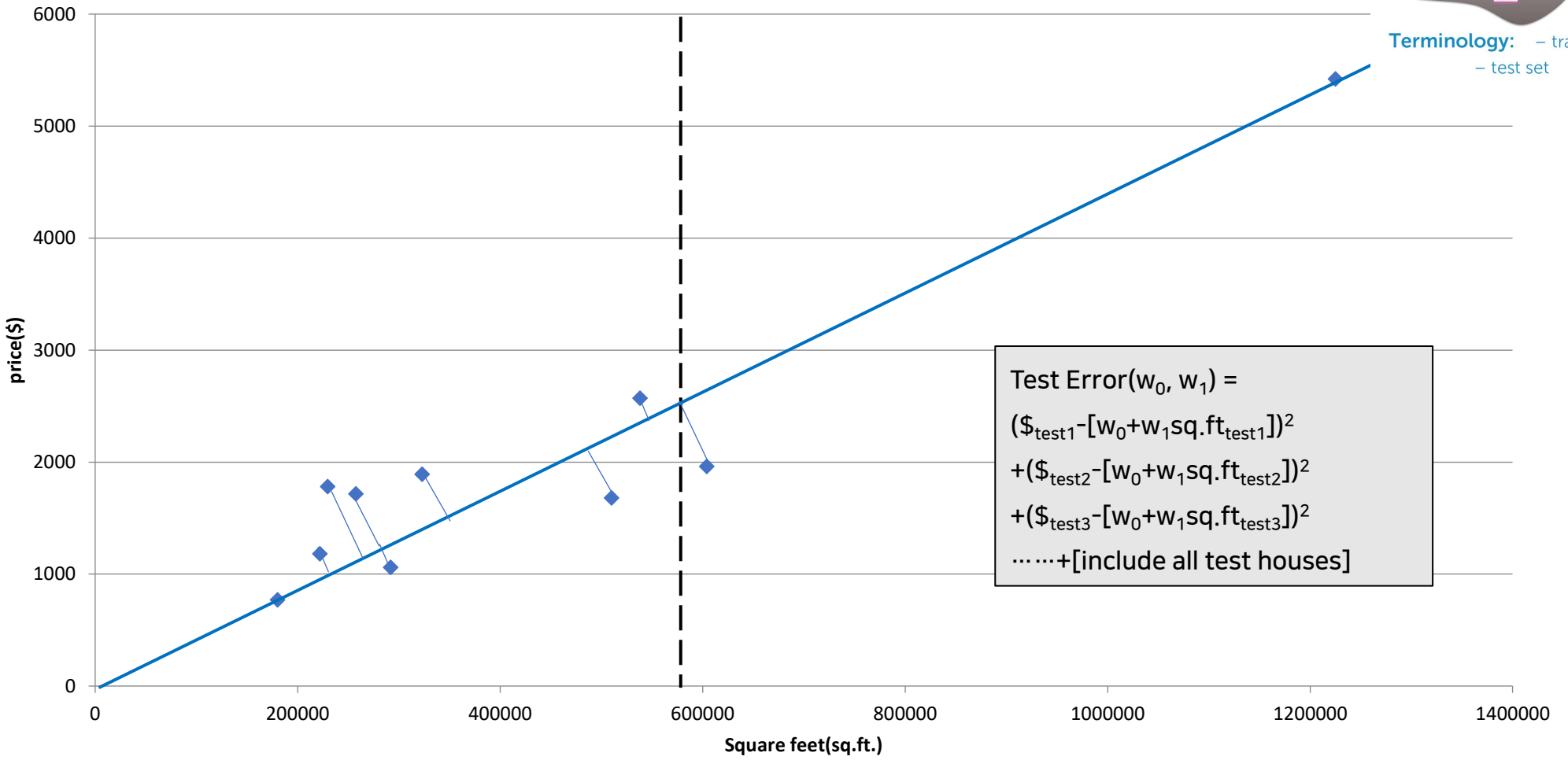
X: Square Feet
Y: Price

RSS = N* MSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

Test Error



Test Error(w_0, w_1) =
($\$_{test1}$-[$w_0$+$w_1$sq.ft$_{test1}$])^2
+($\$_{test2}$-[$w_0$+$w_1$sq.ft$_{test2}$])^2
+($\$_{test3}$-[$w_0$+$w_1$sq.ft$_{test3}$])^2
……+[include all test houses]

Terminology: – training set
– test set

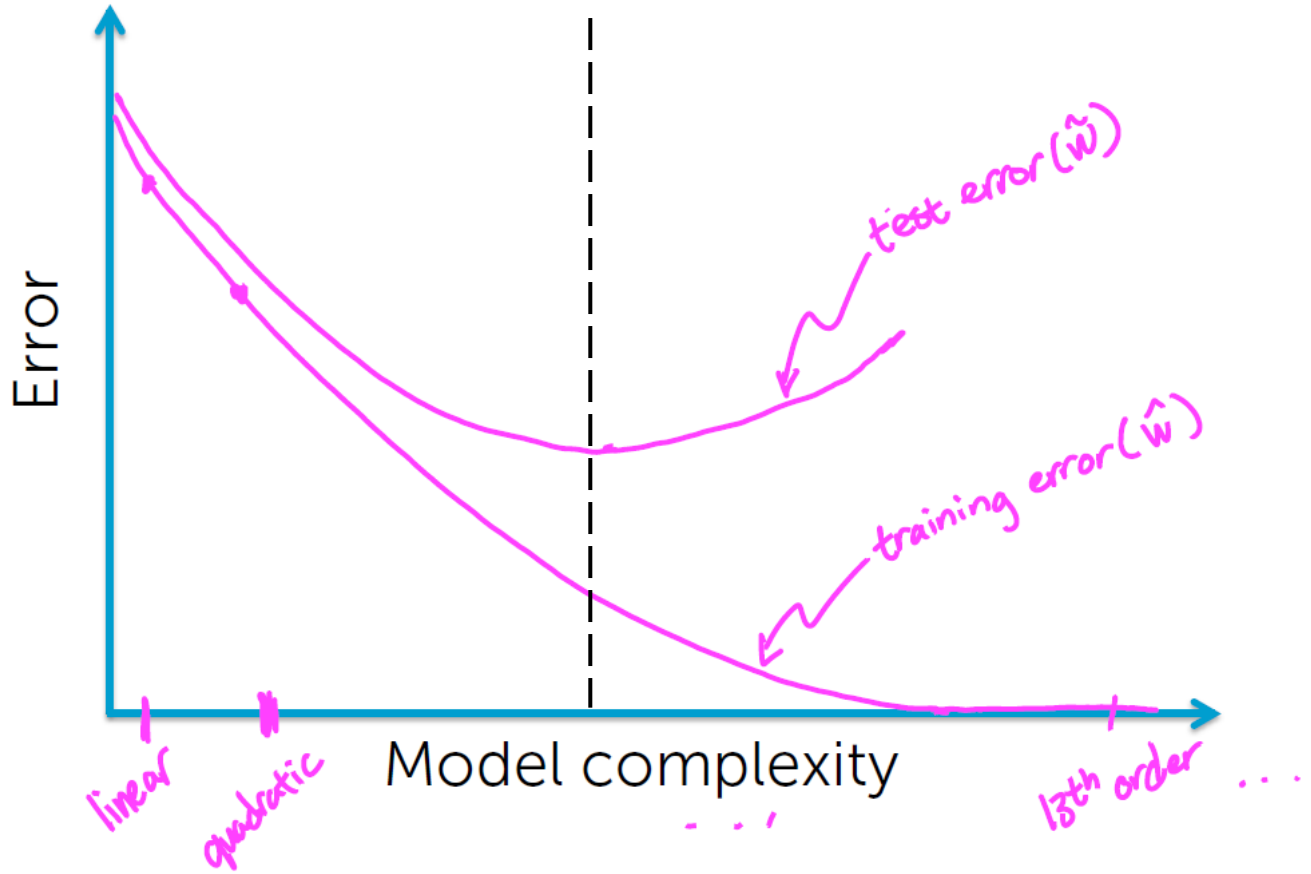
X: Square Feet
Y: Price

RSS = N* MSE

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

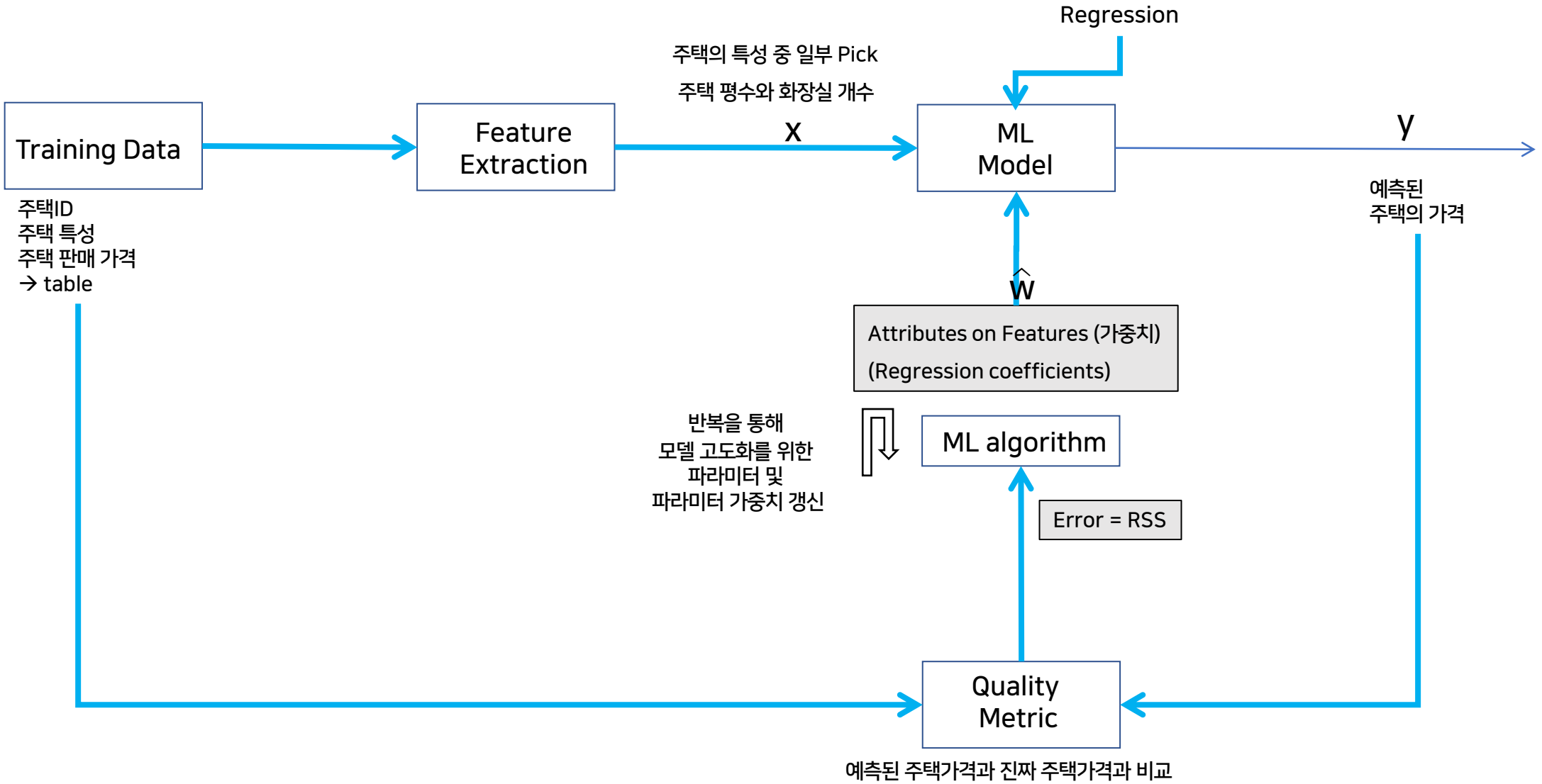
선형회귀(Linear Regression) == 가장 훌륭한 선 긋기, 주어진 데이터들에 가장 근접한 선을 만듭니다.

Training Data/ Test Data 의 Validation



요약하기 : 한 눈에 보기

WorkFlow



	내 용	비 고
1주 (8/29)	강의소개	
2주 (9/5)	Overview	실습준비 자기 소개 section, 조편성 (3인 1조)
3주(9/12)	Regression 이해하기	
4주(9/19)	Regression 실습 및 사례로 배우기	
5주(9/26)	Regression 이해를 점검하기 주제를 선택해서, Regression으로 분석하기 - 주제 선정 - 주제선정이유 (회귀분석이 가능한 이유, 목적 명시) - 분석 - 결과 및 인사이트	Take Home : 중간고사
6주(10/3)	개천절	휴강
7주(10/10)	Midterm Recital	Peer-Review
8주(10/17)	Supervised Learning 이해하기	
9주(10/24)	Supervised Learning / 실습 및 사례로 배우기 Unsupervised Learning 이해하기	
10주(10/31)	Unsupervised Learning 실습하기	
11주 (11/7)	Recommender System 이해하기 / 실습 및 사례로 배우기	
12주(11/14)	프로젝트 설명 및 Proposal 준비시간	프로젝트 설명 및 이전 예제
13주(11/21)	Final Project Proposal	조별 제출 및 발표
14주(11/28)	프로젝트 준비 시간 및 질의/응답시간	
15주(12/5)	Final Presentation	기말고사

Thank You.