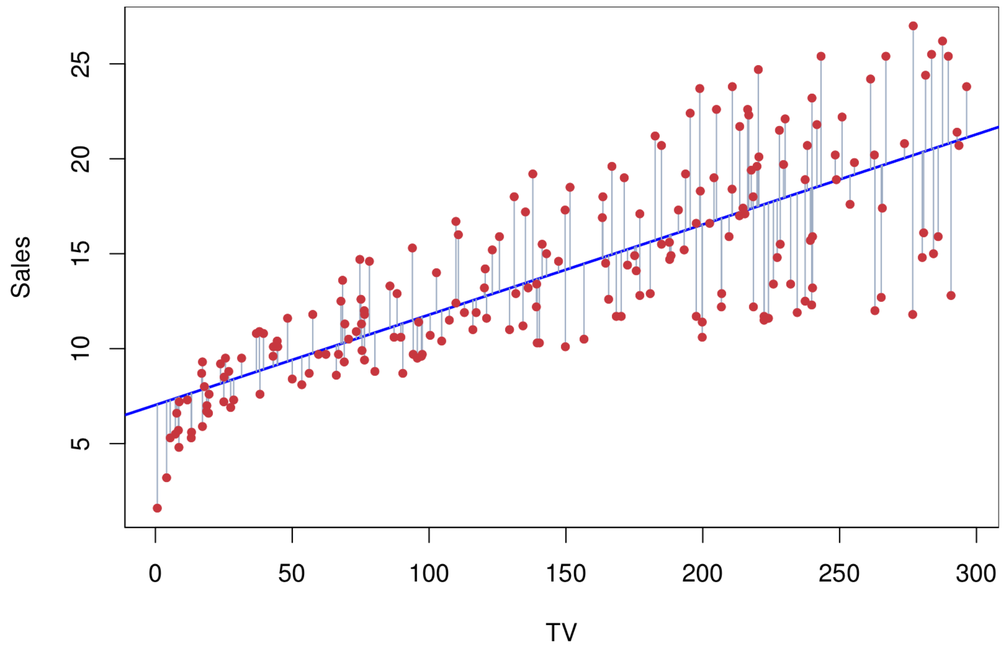

AN
INTRODUCTION
TO
MACHINE
LEARNING
WITH R
DAY 4


DAY 4
Linear Regression

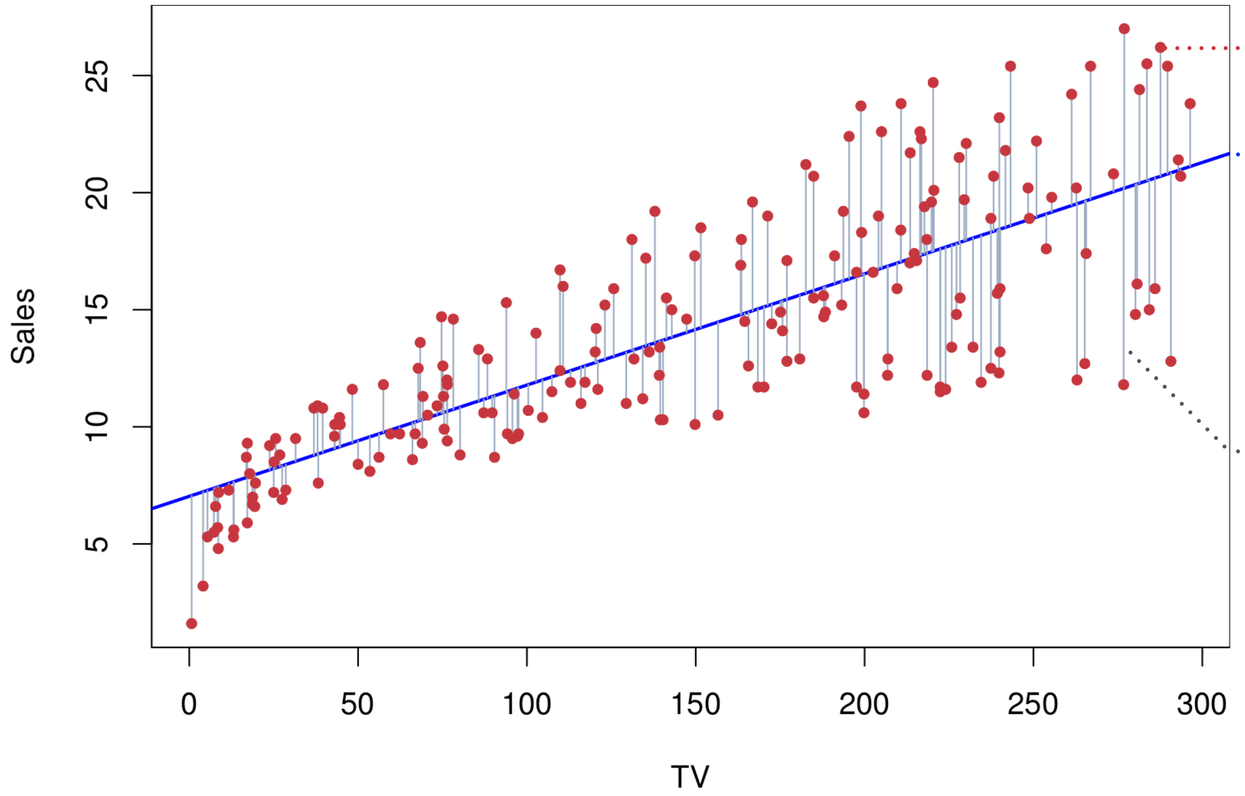
Linear Regression

독립변수와 종속변수 사이의 **관계**를 **선형적**으로 모델링하는 기법



$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Linear Regression

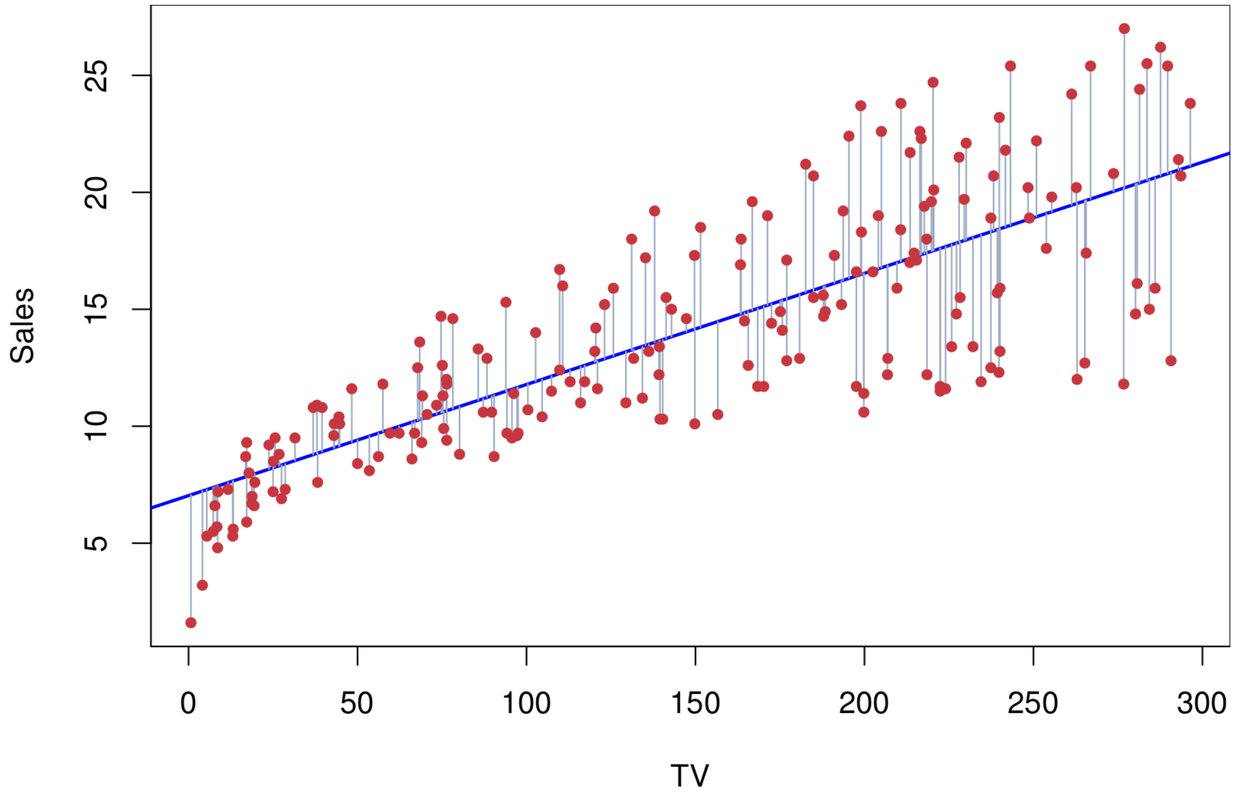


데이터 포인트 (Data Point)

회귀선 (Regression Line)
데이터를 기반으로 알아낸
예측값들을 나타낸 선

잔차 (Residuals)
예측값과 실제값의 차이
$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$

Linear Regression



회귀선은 어떻게 설정될까

잔차제곱합 (Residual Sum of Squares, RSS)

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

Linear Regression

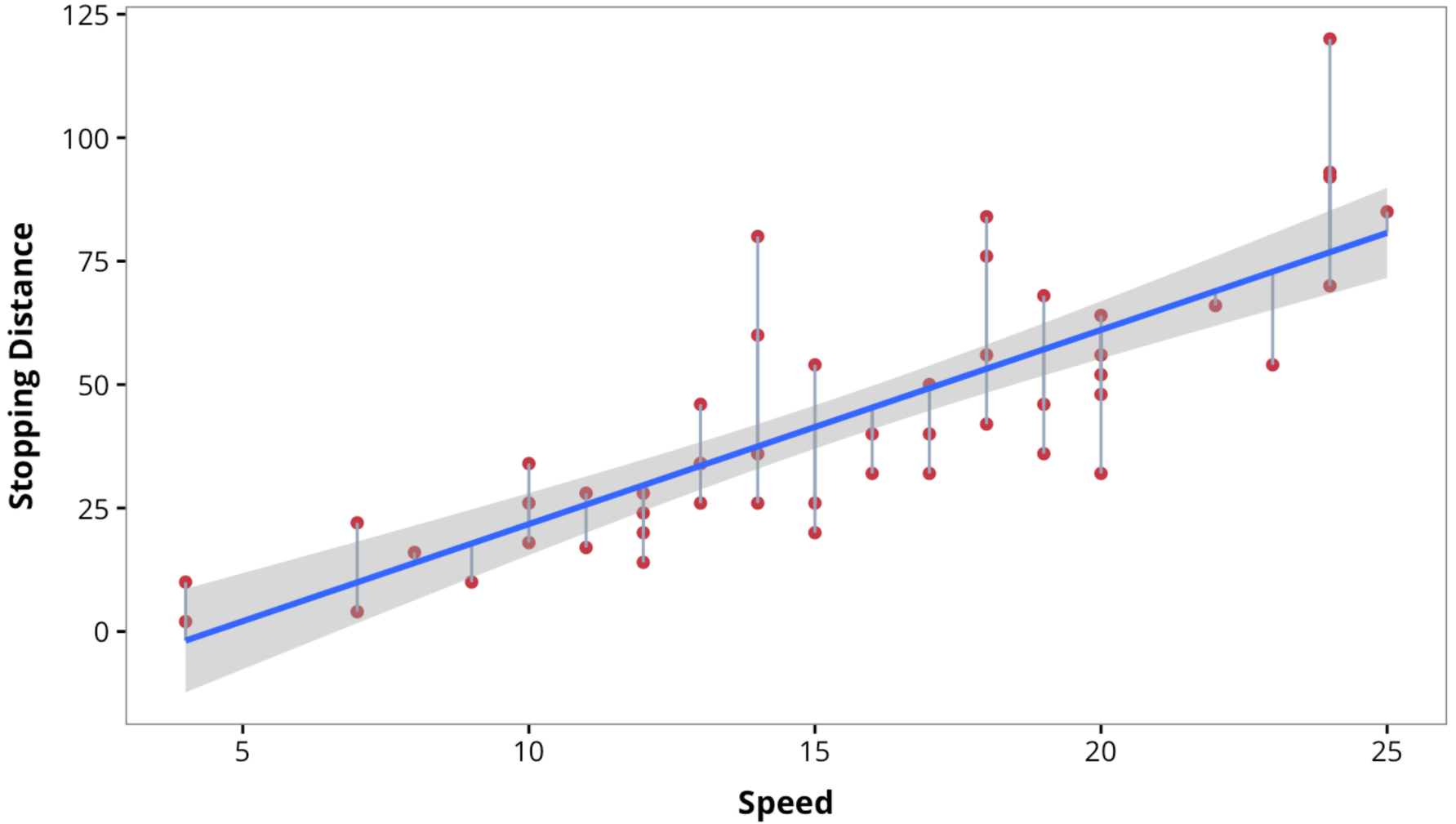
선형회귀법의 기본 가정

- 독립변수 X는 **고정된 값**이다.
- 오차항의 분산이 **동일**하다.
- 오차항간 **상호 독립**이다.
- 오차항의 평균은 **0**이며, 분산은 σ^2인 **정규 분포**를 따른다.
- 독립변수간 **독립**이다.

실습

선형회귀법

STOPPING DISTANCE

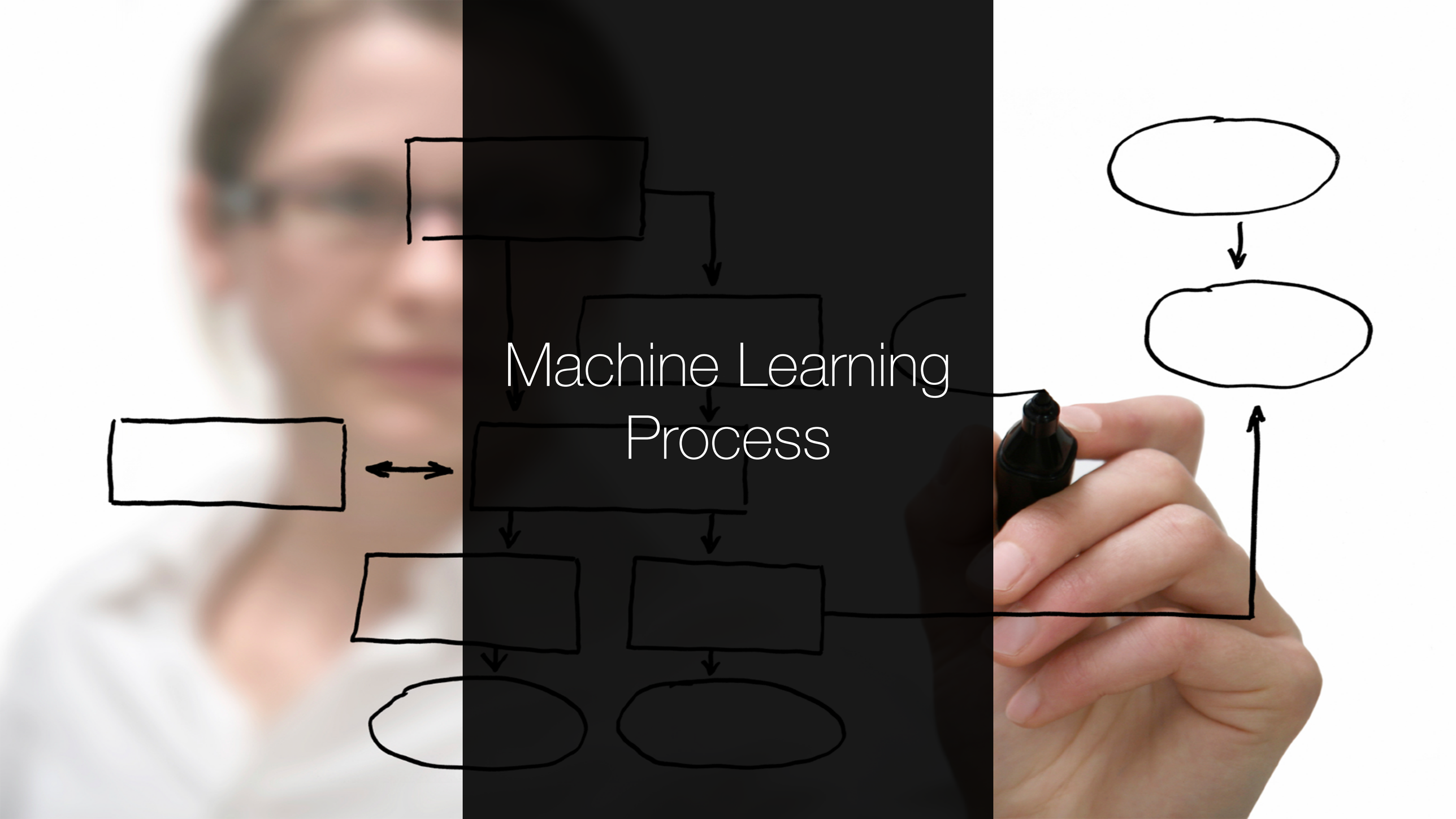

125
100
75
50
25
0
Stopping Distance
5
10
15
20
25
Speed

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
 -29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```


Machine Learning
Process


데이터 수집
Data Collecting
데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing
모델 구축
Building a Model
모델 평가
Model Evaluation
모델 개선
Improving a Model

5%

데이터 수집
Data Collecting

원하는 데이터를 수집한다.

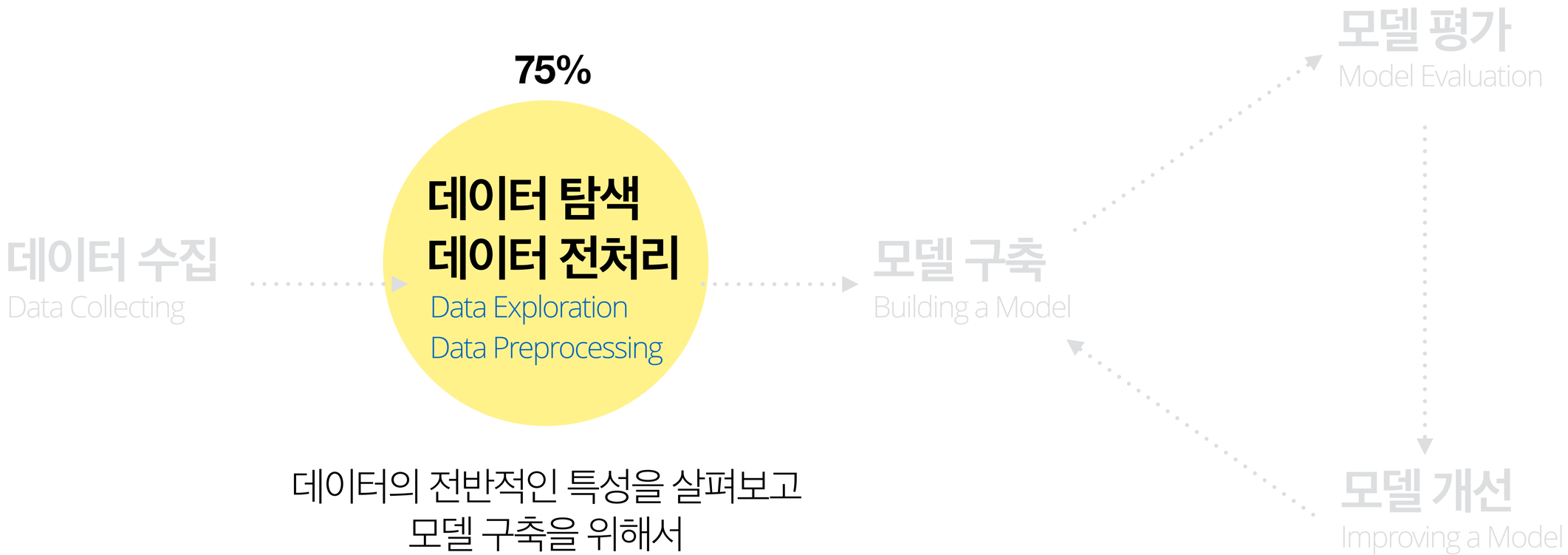
데이터 스크래핑
데이터 크롤링

데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing

모델 구축
Building a Model

모델 평가
Model Evaluation

모델 개선
Improving a Model



데이터의 전반적인 특성을 살펴보고
모델 구축을 위해서
데이터를 형식에 맞춰 수정한다.

데이터 수집
Data Collecting

데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing

5%

모델 구축
Building a Model

모델 평가
Model Evaluation

모델 개선
Improving a Model

전처리한 데이터를 학습하여
원하는 기계학습 모델을 구축한다.

데이터 수집
Data Collecting

데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing

모델 구축
Building a Model

5%

모델 평가
Model Evaluation

구축한 모델을
적절한 메트릭으로 평가한다.

모델 개선
Improving a Model

데이터 수집
Data Collecting

데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing

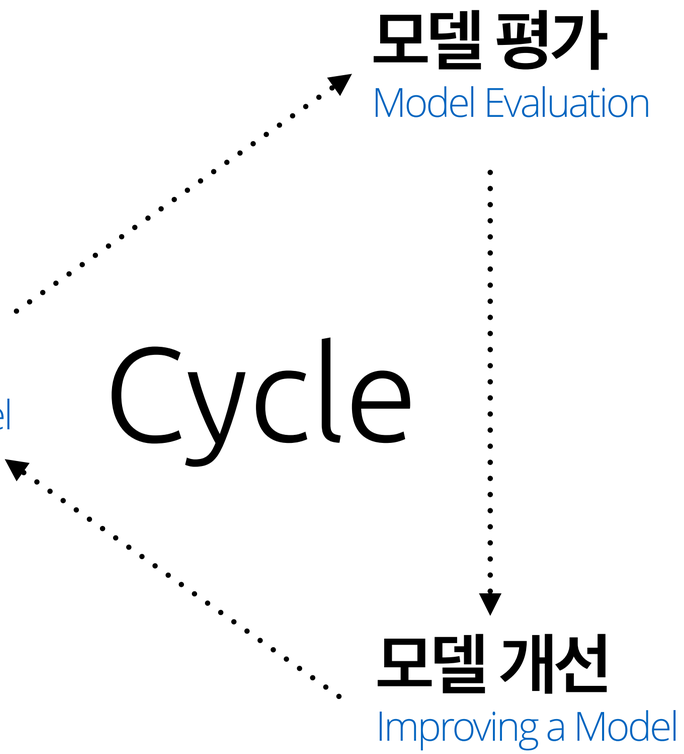
모델 구축
Building a Model

모델 평가
Model Evaluation

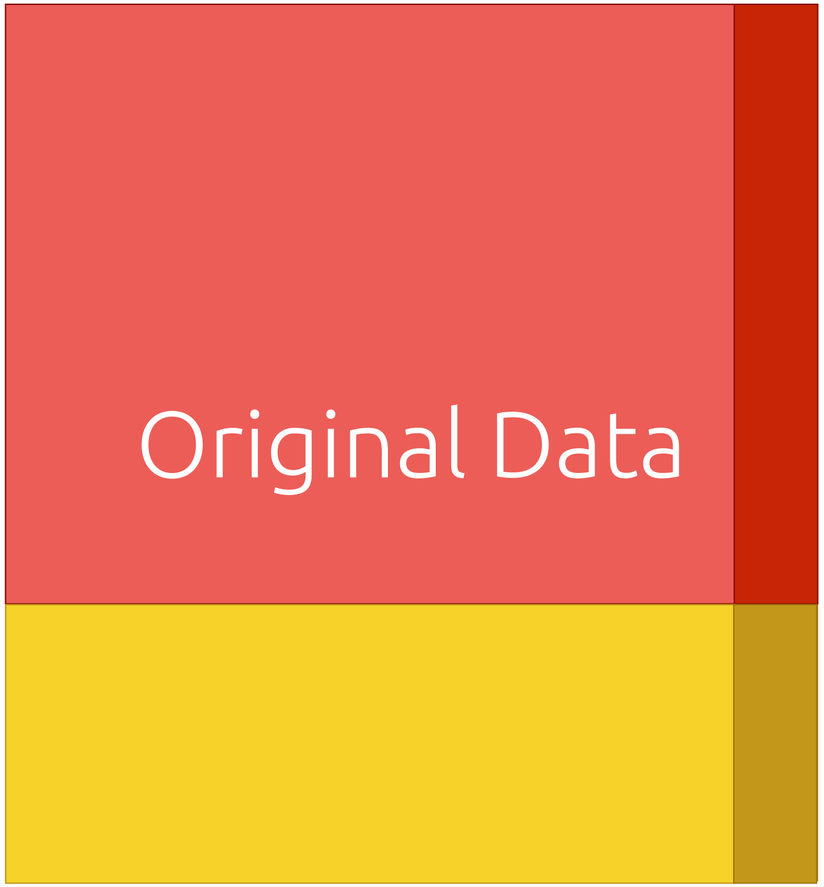
평가를 기준으로
더 나은 모델을 구축한다.

데이터 수집
Data Collecting

데이터 탐색
데이터 전처리
Data Exploration
Data Preprocessing



Split the Data



Test Answer

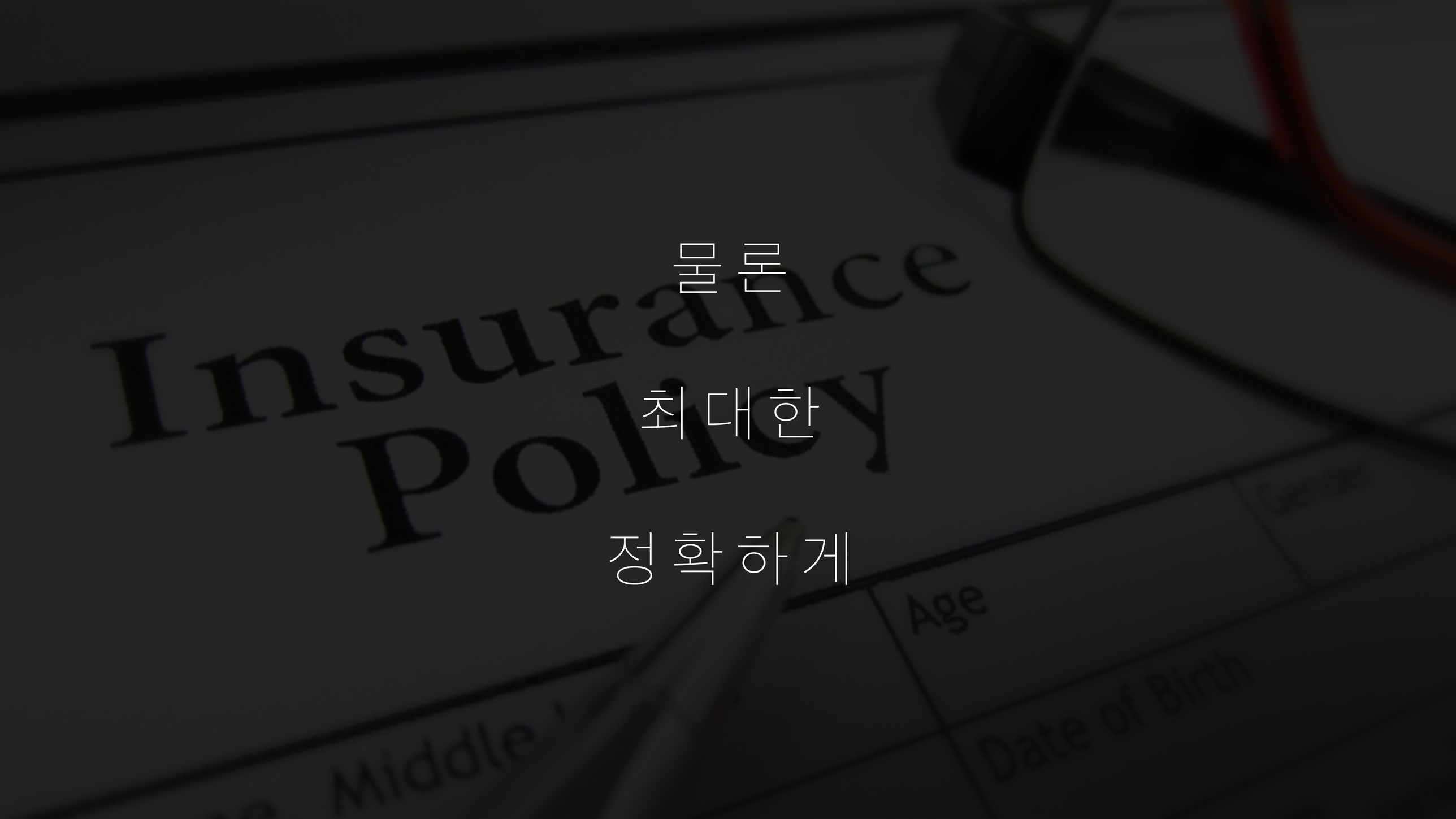
실습

선형회귀법

PREDICT THE CHARGES

가지고 있는 고객의 **데이터**를 활용해서

고객들의 의료비를 **예측**하고자 한다.

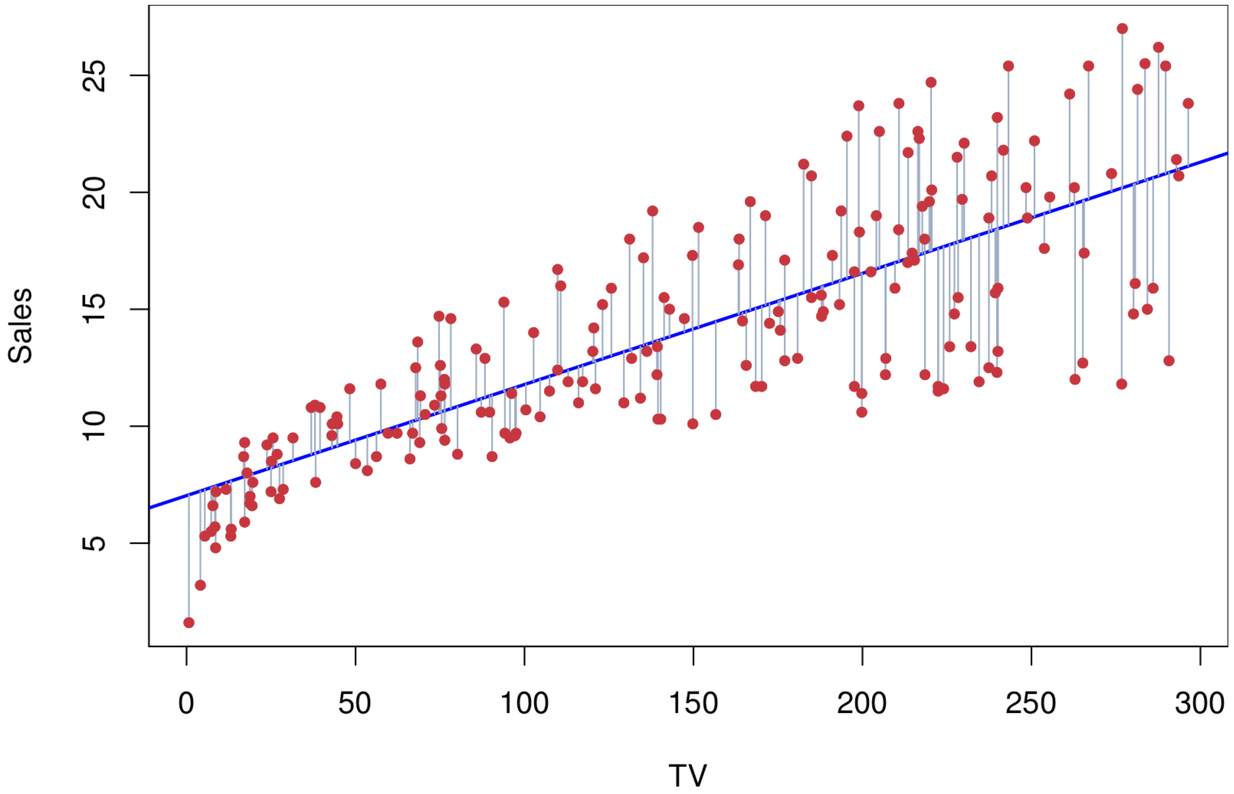


물론

최대한

정확하게

How to Evaluation



회색선을 모두 더해서 평균을 구하자.

평균절대오차 (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

심화

모델 개선하기

[피처 엔지니어링]

FEATURE ENGINEERING

[다중공선성]

MULTICOLLINEARITY

[상호작용항]

INTERACTION TERM

[이상치 제거]

OUTLIER DETECTION

[피처 엔지니어링]

FEATURE ENGINEERING

기존의 변수들에서 예측력에 영향을 미치는
유의미한 변수를 생성하는 과정

또는 모델 구축에 사용할
유의미한 변수들을 선택하는 과정

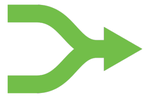
[다중공선성]

MULTICOLLINEARITY

소득

세금

두 변수 사이의 상관관계는 어떻게 될까?
모두 종속변수로 사용해야 할까?



[상호작용항]

INTERACTION TERM

성인병 환자가 있다.

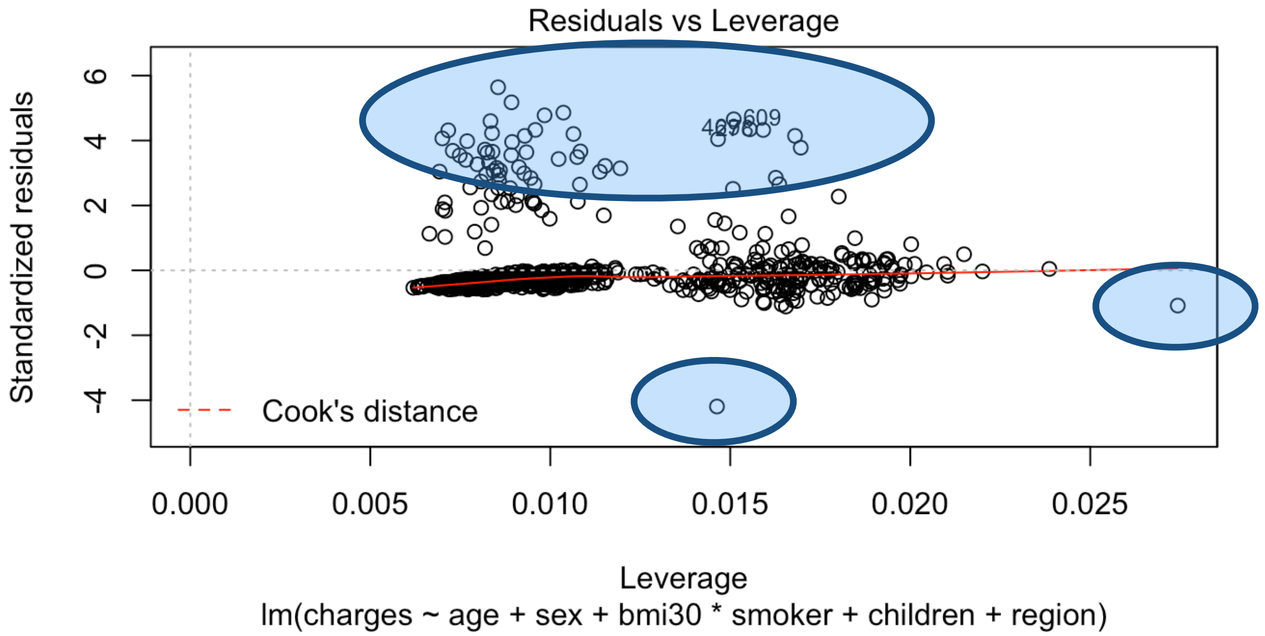
흡연 여부와 비만 여부는 서로 상관이 없다.

그런데

흡연자면서 비만 환자라면?

[이상치 제거]

OUTLIER DETECTION



과제

와인 품질 예측

선형회귀법을 이용해서
와인의 품질을 예측하자

트레이닝 데이터 : `train_wine`
테스트 데이터 : `test_wine`

과제

와인 품질 예측

선형회귀법을 이용해서
와인의 품질을 예측하자

1. **다중공선성**을 가지는 변수들이 있다. 확인 필요!
2. **피처 셀렉션**(Feature Selection)이 필요하다.
3. **이상치**(Outlier)가 많다. 적절히 골라내야 한다.

---

1. **탐색적 데이터 분석**을 통해 인사이트를 얻을 것
2. **MAE**를 이용해서 모델을 평가한다.
3. 구축한 모델을 **설명**할 수 있어야 한다.


THX :)