Linear Regression

고태훈 (taehoonko@dm.snu.ac.kr)

Terminology: 데이터 포인트

❖ 현상을 관측한 단위

- Point (포인트)
- Sample (샘플)
- Instance (인스턴스)
- Record (레코드)
- Observation (관측치)

id	X_1	X_2	…	X_p	Y
1	x_{11}	x_{12}	…	$x_{1,p}$	y_1
2	x_{21}	x_{22}	…	$x_{2,p}$	y_2
…	…	…	…	…	…
n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,p}$	y_n

Terminology: 변수

- 현상들을 설명/표현하는 요소
- Variable, Feature, Attribute, Factor, Field, Column, …

- Predictor variables (예측변수)
- Input variables (입력변수)
- Independent variables (독립변수)

- Target variables (타겟변수)
- Output variables (출력변수)
- Dependent variables (종속변수)

id	X_1	X_2	…	X_p	Y
1	x_{11}	x_{12}	…	$x_{1,p}$	y_1
2	x_{21}	x_{22}	…	$x_{2,p}$	y_2
…	…	…	…	…	…
n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,p}$	y_n

예제: 신용카드회사 고객정보 데이터

- **데이터 포인트: 각 고객 정보**
- **변수: 고객정보를 표현하는 요소**
  - 인구통계학정보: 성별, 생년월일, 나이, 사는 지역, 부양가족 수 등.
  - 신용카드사용내역: 업종 별 결제 내역 및 횟수, 포인트 사용 내역 및 횟수, 신용카드대출 등.
- **$X_1,\ldots X_p$와 Y는 분석 목적에 따라 달라짐**
  - 분석 목적: 고객이 이탈할지 예측
  - Y는 고객들의 이탈 여부 (Yes or No)
  - $X_1,\ldots X_p$는 고객을 설명하는 변수들: 위에 언급한 인구통계학정보, 신용카드사용내역 등.

다중선형회귀 (Multiple linear regression)

❖ 목표

- 수치형 출력변수 Y를 여러 개의 입력변수 X_1, X_2, ..., X_p의 선형조합으로 표현하는 식을 도출하는 것

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X_1} + \beta_2 \mathbf{X_2} + \ldots + \beta_p \mathbf{X_p} + \boldsymbol{\varepsilon}$$

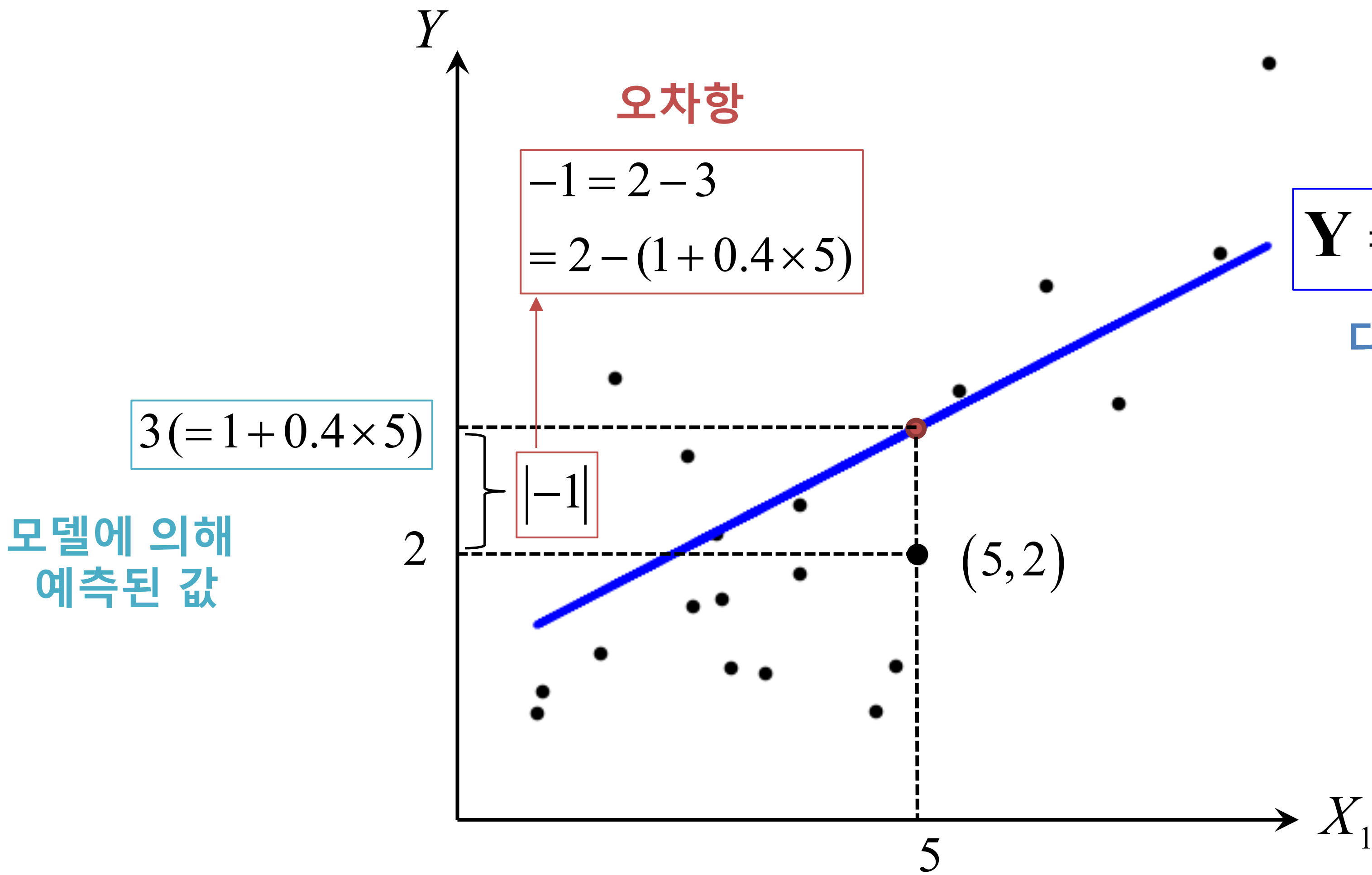
Intercepts
(절편)

coefficients
(계수)

vector of error
variables
(오차항으로
이루어진 벡터)

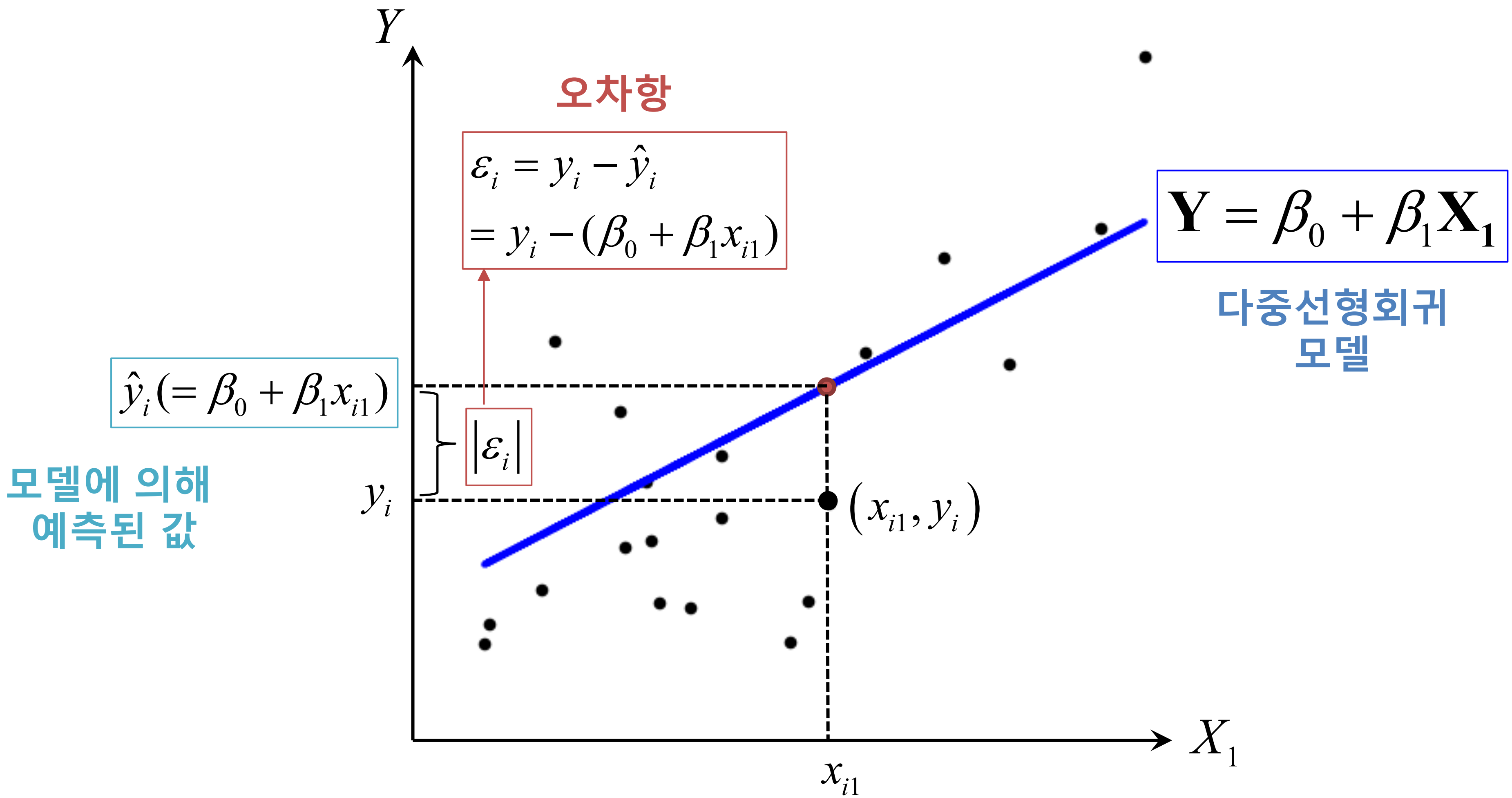
다중선형회귀

❖ 다중선형회귀모델의 계수들을 알고 있다고 가정했을 때,



다중선형회귀

❖ 다중선형회귀모델의 계수들을 알고 있다고 가정했을 때,



다중선형회귀

❖ 다중선형회귀모델의 계수들을 알고 있다고 가정했을 때,

$$\hat{\mathbf{Y}} = \beta_0 + \beta_1 \mathbf{X}_1 \qquad \hat{y}_i = \beta_0 + \beta_1 x_{i1}$$

id	X_1	Y
1	x_{11}	y_1
2	x_{21}	y_2
…	…	…
i	x_{i1}	y_i
…	…	…
n	$x_{n,1}$	y_n

$\hat{Y}$	$\boldsymbol{\varepsilon}$
$\hat{y}_1$	ε_1
$\hat{y}_2$	ε_2
…	…
$\hat{y}_i$	ε_i
…	…
$\hat{y}_n$	ε_n

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} \qquad \varepsilon_i = y_i - \hat{y}_i$$

다중선형회귀

❖ 다중선형회귀모델의 계수들을 알고 있다고 가정했을 때,

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

id	X_1	X_2	…	X_p	Y
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{21}	…	x_{2p}	y_2
…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ip}	y_i
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}	y_n

$\hat{\mathbf{Y}}$	$\boldsymbol{\varepsilon}$
$\hat{y}_1$	ε_1
$\hat{y}_2$	ε_2
…	…
$\hat{y}_i$	ε_i
…	…
$\hat{y}_n$	ε_n

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$\varepsilon_i = y_i - \hat{y}_i$$

다중회귀분석모델

❖ In matrix form,

- **X**: n by (p+1) matrix / **y**: n by 1 vector / **β**: p by 1 vector

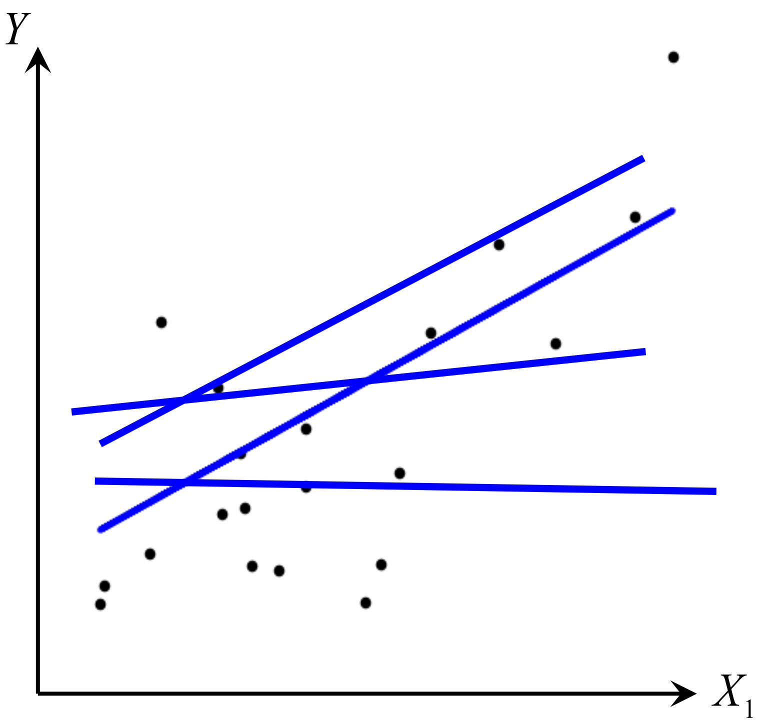
$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \ldots + \beta_p\mathbf{X_p} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$

다중선형회귀

❖ 실제로는,

- 데이터가 주어진 상태이며, 다중선형회귀모델의 계수는 모름.
- 즉, 어떠한 회귀모델이 현 데이터에 더 적합한지 모르는 상태



Which regression model is the best?

In other words,
which coefficient set $\boldsymbol{\beta}$ is the best?

다중선형회귀

❖ 다중선형회귀모델의 계수들을 모를 때,

- 추정하고자 하는 계수들을 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$라 하자.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

id	X_1	X_2	…	X_p	Y
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{21}	…	x_{2p}	y_2
…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ip}	y_i
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}	y_n

$\hat{Y}$	$\boldsymbol{\varepsilon}$
$\hat{y}_1$	ε_1
$\hat{y}_2$	ε_2
…	…
$\hat{y}_i$	ε_i
…	…
$\hat{y}_n$	ε_n

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$\varepsilon_i = y_i - \hat{y}_i$$

다중선형회귀모델의 학습 = 계수추정

❖ 따라서,

- 주어진 데이터를 이용하여 선형회귀모델의 계수를 **추정(estimation)** 해야 한다.

❖ 추정 방법 중 하나인 Ordinary least squares (OLS)

- 가장 단순한 추정 방법
- **<u>오차의 제곱합</u>**을 **최소화**하는 계수 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$를 찾는 것

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left\{ y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \right) \right\}^2$$

Ordinary least squares (OLS)

❖ **목적식(오차제곱합)을 각 계수로 편미분하여 계수들을 도출**

- 한 개의 독립변수만을 이용한 회귀분석모델의 경우,

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left\{ y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \right) \right\}^2$$

$$\frac{\partial}{\partial \beta_0} \left(\sum_{i=1}^{n} \varepsilon_i^{\,2} \right) = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) = 0$$

$$\frac{\partial}{\partial \beta_1} \left(\sum_{i=1}^{n} \varepsilon_i^{\,2} \right) = \sum_{i=1}^{n} -2x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) = 0$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_{i1} - \overline{x})^2}$$

Ordinary least squares (OLS)

- In matrix form,
  - $\mathbf{X}$: n by (p+1) matrix / $\mathbf{Y}$: n by 1 vector / $\hat{\boldsymbol{\beta}}$: p by 1 vector

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = -2(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X} = \mathbf{0}$$

$$-\mathbf{Y}^{\mathrm{T}}\mathbf{X} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\boxed{\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\end{aligned}}$$