

## II. 일반 회귀분석 (Regression)

### ① Linear Regression I

# Supervised Learning

## Regression

Linear Regression

Ordinary Least Squares  
Regression

LOESS (Local Regression)

Neural Networks

## Classification

Decision Trees

Support Vector Machine

Naïve Bayes

K-Nearest Neighbours

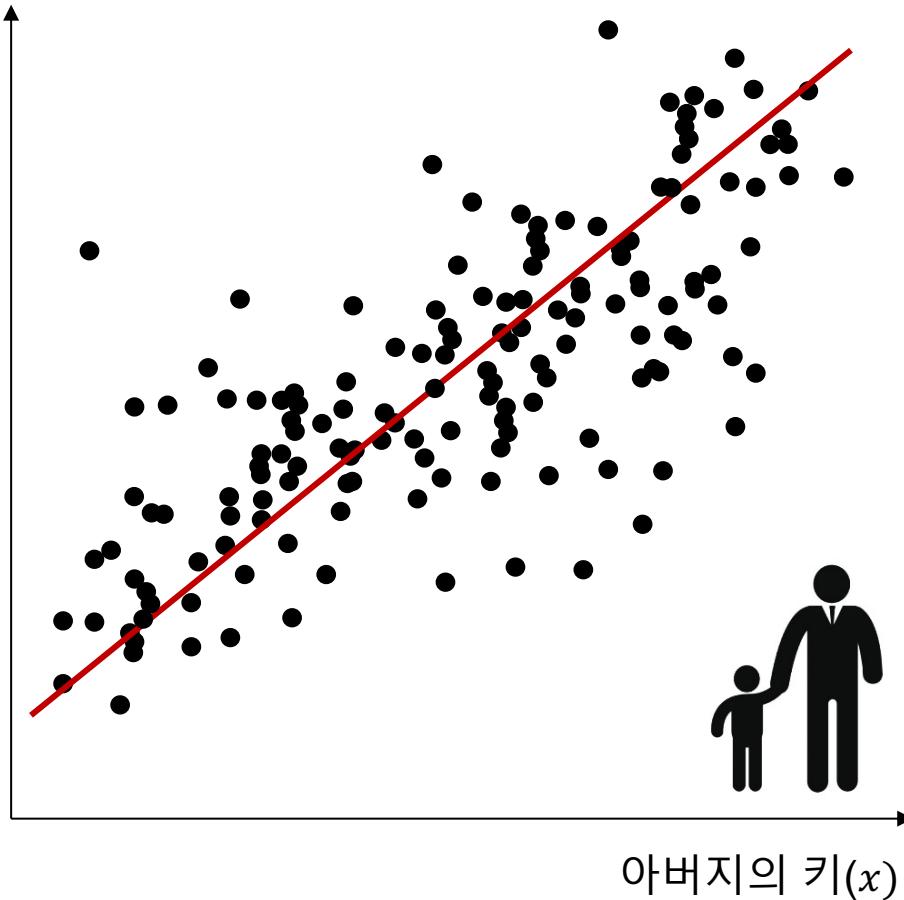
Logistic Regression

Random Forests

# 회귀분석(Regression)이란 무엇인가?

회귀분석은 인과관계(Causality)를 확인하기 위해 고안된 모델이며, 1885년 유전학자인 프란시스 골튼이 부모의 키와 자녀의 키 사이에 관계가 있는지 파악하기 위한 연구에서 시작됨

아들의 키(y)



아버지의 키와 아들의 키 사이에 선형적 관계가 있음.  
키가 더 커지거나 키가 더 작아지는 것이 아니라 주어진 아버지의 키 안에서 아들의 키 평균으로 돌아가려는 성향이 있으며, 이를 [평균에 회귀\(Regress\)한다고 표현함](#). 즉, 아버지의 키가 주어졌을 때, 아들의 조건부 키 평균들을 이으면 이것이 회귀선(Regression line)이 됨

선형회귀분석은 데이터의 선형관계(Linear Relationship)을 가장 잘 설명하는 선(line)을 그은 것

관계의 가정에 따라

- 선형회귀분석(Linear Regression)
- 비선형 회귀분석(Non-linear Regression)

변수의 수에 따라

- 단순 회귀분석(Simple Regression : 독립변수 1개)
- 다중 회귀분석(Non-linear Regression) : 독립변수 2개 이상

# 회귀분석의 적용 및 응용

---

“광고가 매출액에 얼마나 영향을 미치는가?”

“제품판매 가격을 어떻게 결정할 것인가?”

“매장을 오픈했을 때, 예상되는 매출은 얼마일까?”

“소비자 수요가 얼마나 될까?”

:

“비만이 성인병 발생에 미치는 영향은 얼마나 될까?”

“세포활동이 암 발병에 미치는 영향은 어떻게 될까?”

“신약의 성분변화에 따라 치료율이 얼마나 될까?”

:

“교육수준에 따라 소득이 어떻게 달라질까?”

“조직성과가 직무만족에 미치는 영향은 얼마나 될까?”

:

회귀분석은 인과관계의 정도를 정량적으로 분석하고자 하는 여러 맥락에서 적용가능함

# 선형 회귀분석(Linear Regression)

선형 회귀분석(Linear regression)은 우리가 모르는 어떤 인과의 사회현상을 선형관계를 가정하고, 관계를 추정하는 것임

## Regression Model의 수식 표현

$$y_i = \underline{\alpha + \beta x_i} + \underline{\varepsilon_i}$$

내가 가정한 관계에서  
회귀모형으로  
설명되는 부분

**오차항(Error term)**  
: 회귀모형으로  
설명되지 않는 부분

### God's Model

우리가 모르는 모집단에서의 ' $x, y$  관계가 이럴 것이다'  
가정한 식. **오직 신(God)만 알고 있음**

$\alpha, \beta, \varepsilon$ 는 모르지만 모집단의 관계를 설명해주는 모수  
(Parameter)로 추정의 대상이 됨.

$$y_i = \underline{\hat{\alpha} + \hat{\beta} x_i} + \underline{e_i}$$

실제 Data에서  
회귀모형으로  
설명되는 부분

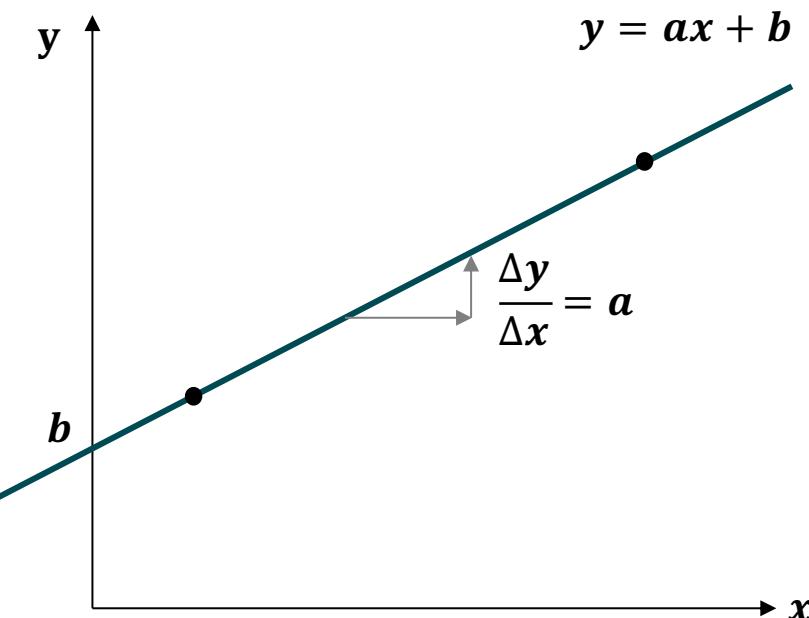
**잔차(Residual)**  
: 회귀모형으로  
설명되지 않는 부분

### 나의 Model

실제 Data로 관계를 추정한 식  
나의 모델에서 모수인  $\beta$ 를 추정한 기울기(Slope)  
를 회귀계수(Coeffieicnt)라고 하고, 모수와의 표현  
상 구분을 위해 햇(Hat) 표시를 함

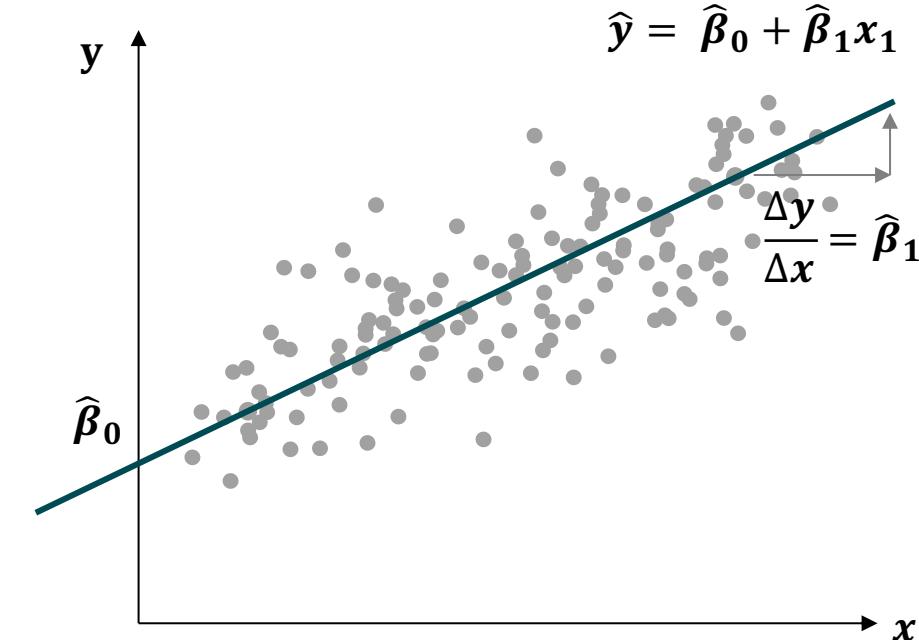
# 선형 회귀분석과 1차 방정식

Remind 1<sup>st</sup> year of mid school...



- ✓ 기울기(Slope)와 절편(Intercept)이 주어지면 1차방정식 즉, 1차 함수(선형함수) 형태로 나타낼 수 있음
- ✓ 1차 방정식에서는 기울기와 절편은 주어지는 값

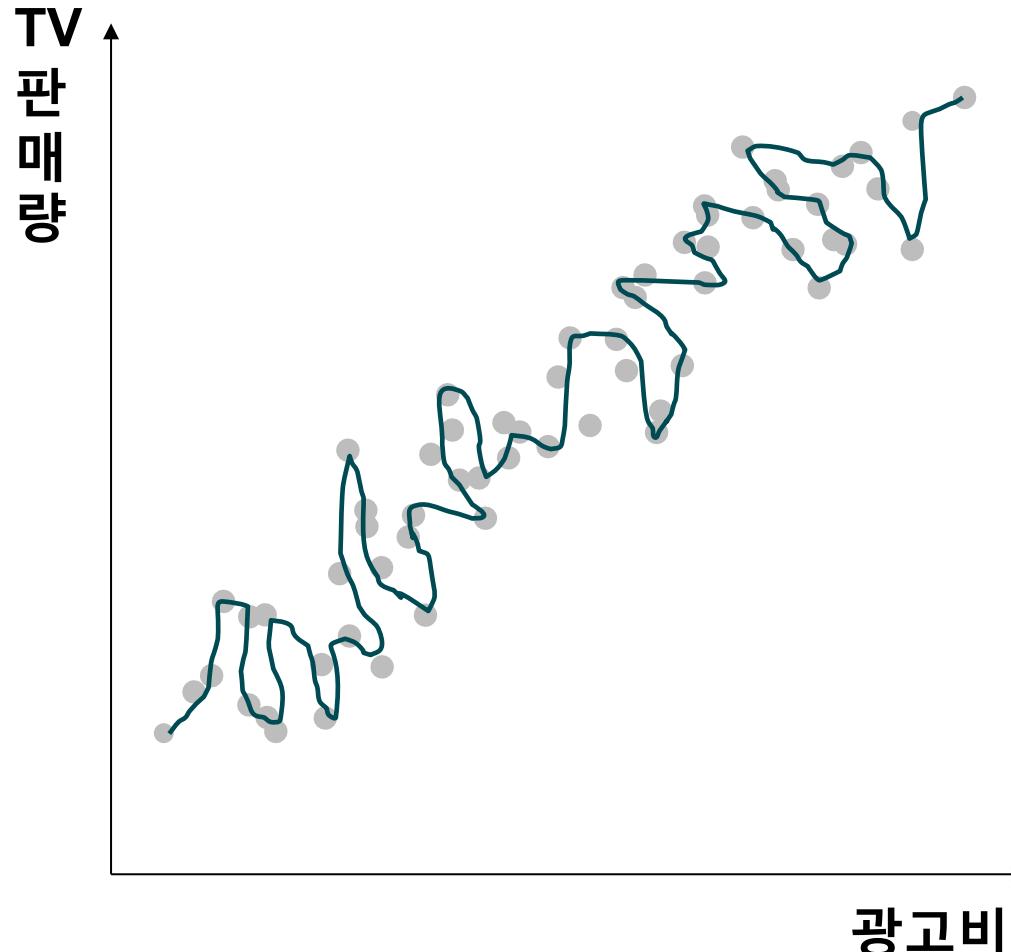
회귀분석 = 1차방정식을 찾는 과정



- ✓ 회귀분석은 1차 함수로 관계를 **가장 잘** 나타내기 위해 데이터로부터 역으로 기울기(Slope)와 절편(Intercept)을 찾아가는 과정
- ✓ 기울기의 크기에 따라 두 변수 간 인과관계의 민감도를 판단할 수 있음

# 과적합 문제(Overfitting problem)

왜 그럼 오차(Error)를 허용하면서 직선으로 추정하는 것인가? 모든 점을 지나도록 곡선으로 추정하면 안되는 것인가? 다음의 예시를 보자.



## 과적합 문제(Overfitting Problem)

- ✓ 모형을 학습시키는 목적은 표본으로부터 일반적인 결과를 도출해 모집단의 특성을 추정하고, 예측하는 것
- ✓ 과적합의 문제가 생기면, 모형의 적합도는 높으나 새로운 자료(Data)에 대한 예측 성능이 좋지 않음
- ✓ 우리 목적은 Training을 잘하는 것도 중요하지만 Prediction을 잘하는 것이 더욱 중요함
- ✓ 따라서, 선형 모형의 목적은 오차를 어느정도 허용하더라도 선형관계를 통해 예측성능을 높이고자 하는 것임

***“Linear is beautiful”***

# 회귀식 수립과 회귀계수의 해석

---

회귀식

$$y_i = \boxed{350} + \boxed{250x_{1i}} - \boxed{450x_{2i}} + e_i$$

상수항(Constant)      계수(Coefficient)      잔차(Residual)

해석할 땐 제외!

회귀계수 해석

독립변수  $x$ 들이  
모두 0일 때의 값

독립변수  $x_1$ 이 1변할 때,  
 $y$ 의 변화량      독립변수  $x_2$ 이 1변할 때,  
 $y$ 의 변화량

## **선형회귀분석 예시**

# 회귀분석 예시#1

---

OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계



GDP 

A 3D rendering of the word "GDP" in black letters on white cubes. To the right of the letter "P", there is a green cube with a white upward-pointing arrow icon.

# 회귀분석 예시#1

## OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계

```
call:  
lm(formula = gdp ~ selfemp, data = selfemp)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-35092 -9593     390    8657   47019  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 67588     7153    9.449 2.35e-10 ***  
selfemp     -1682      395   -4.259 0.000197 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 17320 on 29 degrees of freedom  
Multiple R-squared:  0.3848,    Adjusted R-squared:  0.3636  
F-statistic: 18.14 on 1 and 29 DF,  p-value: 0.0001972
```

절편  
기울기

모형의 설명력

# 회귀분석 예시#1

## OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계

Dependent Variable : gdp				
Variables	$\hat{\beta}$	Std. Err	t	P >  t
Selfemp	-1682.166	394.9608	-4.26	0.000
(intercept)	67587.9	7153.121	9.45	0.000
R-squared : 0.3848 / Adjusted R-squared : 0.3636				

### 1. 모델 적합도 설명

-  $R^2$  값이 0.385로 모형 설명력은 38.5% 정도인 것으로 나타남

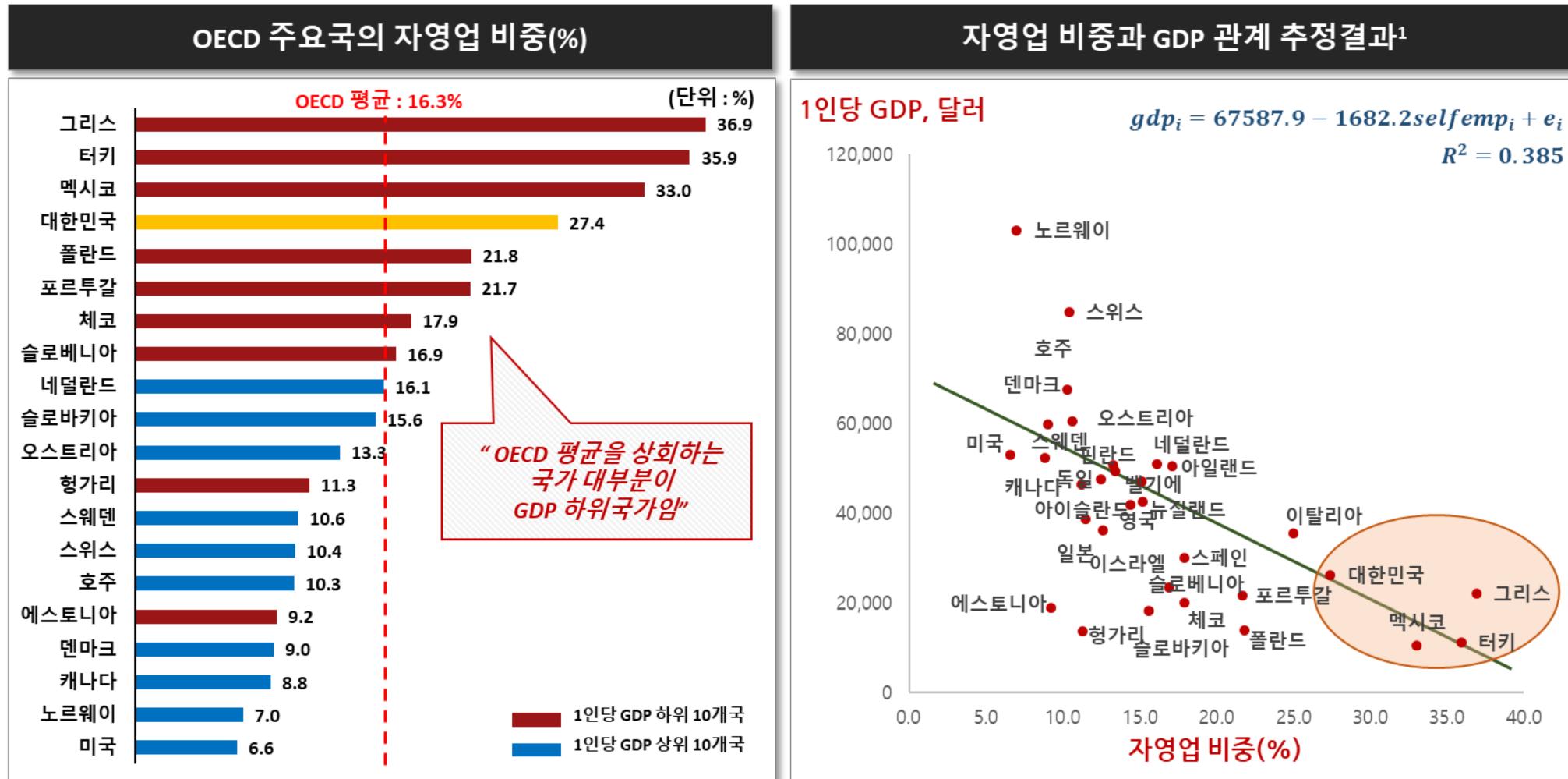
### 2. 계수의 설명

$$\text{회귀식 : } gdp_i = 67587.9 - 1682.2selfemp_i + e_i$$

자영업 비중이 1% 증가하면, 1인당 GDP가 -1682.2 달러(\$) 감소할 것으로 예상할 수 있음

# 회귀분석 예시#1

## OECD 주요 국가의 1인당 GDP와 자영업자 비중 간 관계



1. 자영업자 비율과 1인당GDP 선형회귀모형 추정결과로, 모형 설명력은 38.5%임  
Source: OECD stat. 2013.

## 회귀분석 예시#2

도요타(Toyota) 코롤라(Corolla) 중고차 가격 결정 모형



# 회귀분석 예시#2

## 도요타(Toyota) 코롤라(Corolla) 중고차 가격 결정 모형

```
call:  
lm(formula = Price ~ ., data = toyota_train)  
  
Residuals:  
    Min      1Q   Median      3Q     Max  
-10775.1  -744.9   -27.1   751.5  6362.3  
  
Coefficients:  
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.049e+04 1.656e+03 -6.334 3.61e-10 ***  
Age          -1.156e+02 2.974e+00 -38.859 < 2e-16 ***  
KM           -1.678e-02 1.521e-03 -11.032 < 2e-16 ***  
FuelTypeDiesel 3.201e+03 6.061e+02  5.281 1.58e-07 ***  
FuelTypePetrol 1.833e+03 4.154e+02  4.413 1.13e-05 ***  
HP           5.204e+01 6.231e+00  8.351 2.26e-16 ***  
MetColor      2.459e+02 8.583e+01  2.865 0.00426 **  
Automatic     1.747e+02 1.791e+02  0.975 0.32978  
CC            -3.470e+00 6.036e-01 -5.749 1.19e-08 ***  
Doors         -1.085e+02 4.652e+01 -2.333 0.01982 *  
weight        2.545e+01 1.551e+00 16.405 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1261 on 996 degrees of freedom  
Multiple R-squared:  0.8776,    Adjusted R-squared:  0.8764  
F-statistic: 714.4 on 10 and 996 DF,  p-value: < 2.2e-16
```

절편

기울기

# 회귀분석 예시#2

## 도요타(Toyota) 코롤라(Corolla) 중고차 가격 결정 모형

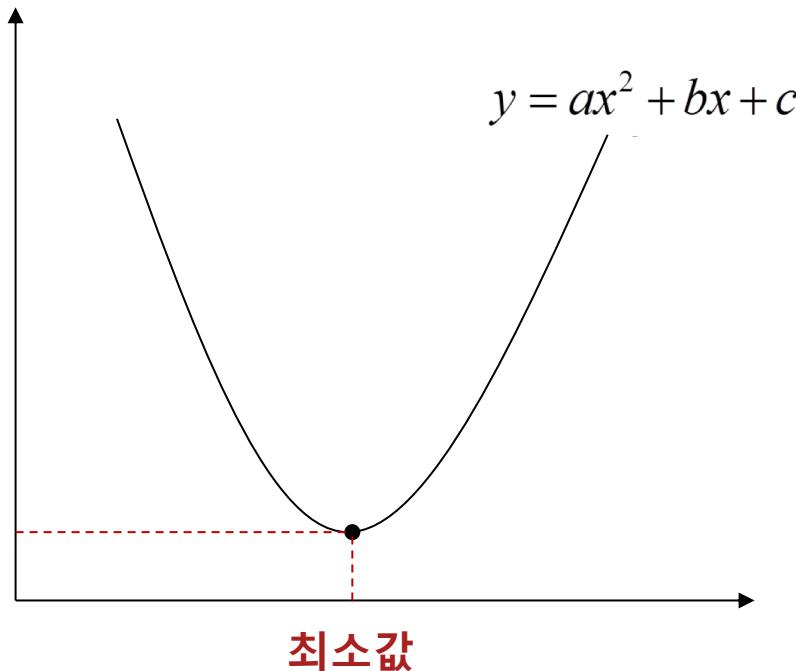
Dependent Variable : Price				
Variables	$\hat{\beta}$	Std. Err	t	P >  t
Age	-115.60	2.97	-38.859	< 2e-16
KM	-0.02	0.00	-11.032	< 2e-16
FuelTypeDiesel	3201.00	606.10	5.281	1.58E-07
FuelTypePetrol	1833.00	415.40	4.413	1.13E-05
HP	52.04	6.23	8.351	2.26E-16
MetColor	245.90	85.83	2.865	0.00426
Automatic	174.70	179.10	0.975	0.32978
CC	-3.47	0.60	-5.749	1.19E-08
Doors	-108.50	46.52	-2.333	0.01982
Weight	25.45	1.55	16.405	< 2e-16
(intercept)	-10490.00	1656.00	-6.334	3.61E-10

$$\begin{aligned} \text{Price} = & -10,490 - 115.60\text{AGE} - 0.02\text{KM} + 3,201.00\text{FuelTypeDiesel} + 1,833.00\text{FuelTypePetrol} \\ & + 52.04\text{HP} + 245.90\text{MetColor} + 174.70\text{Automatic} - 3.47\text{CC} - 108.50\text{Doors} + 25.45\text{Weight} \end{aligned}$$

# 참고 - 회귀분석의 기울기와 절편은 어떻게 찾을까?

2차 함수의 미분

Ex) risk, error, loss



$$y = ax^2 + bx + c$$

↓ 미분

$$y' = 2ax + b$$

$$y'_{x=0} = 2a \times 0 + b = b$$

$\frac{\text{---}}{\text{---}}$

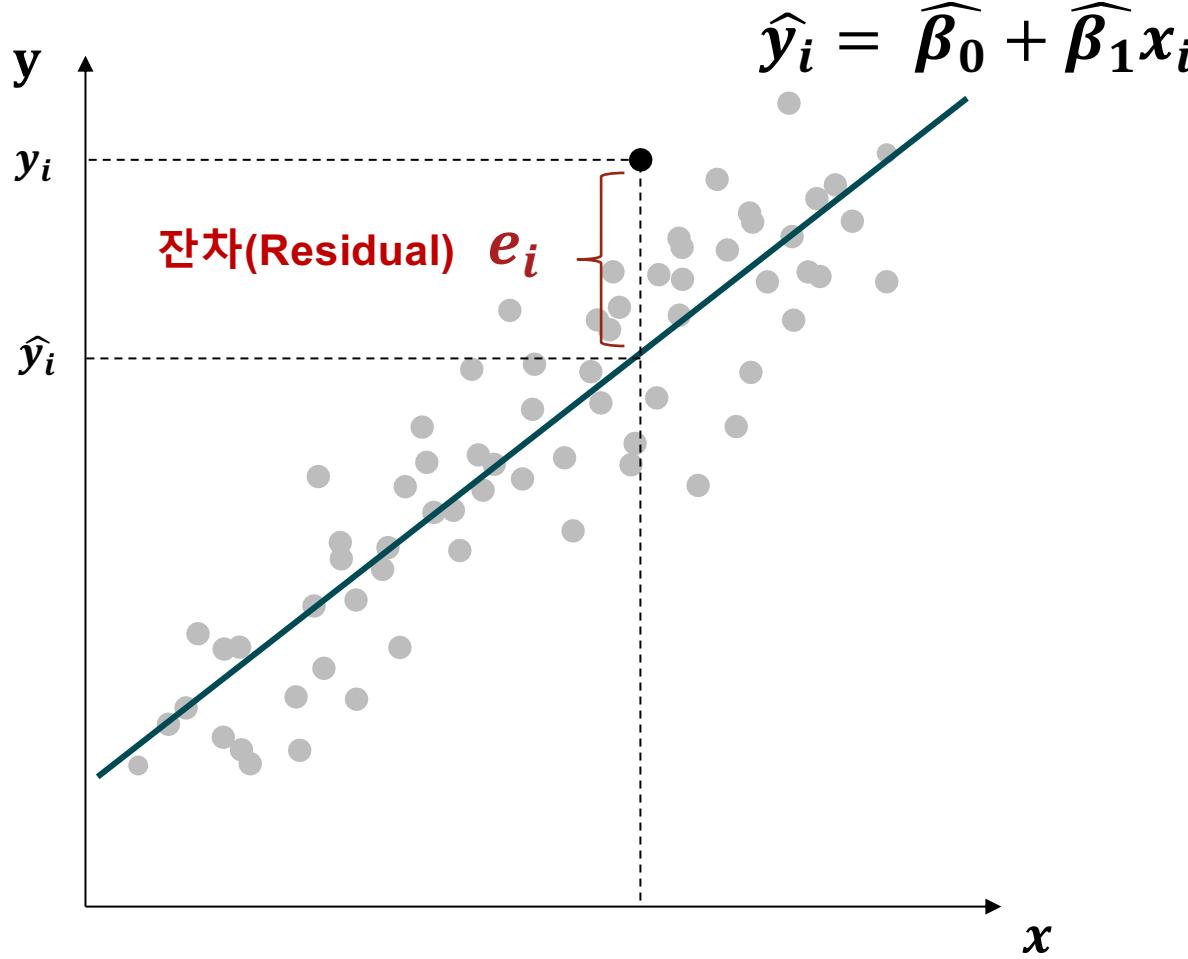
$x=0$ 에서 미분계수  
 $\Rightarrow x=0$ 에서 접선의 기울기  
 $\Rightarrow$ 즉  $y$ 절편의 접선의 기울기

만약, Error를 2차함수의 형태로 나타낸다면, 미분을 통해 회귀분석에서 Error를 최소화하는 기울기(slope)와 절편(intercept)을 구할 수 있음!

## 참고 - 회귀분석의 기울기와 절편은 어떻게 찾을까?

---

최소제곱법(OLS : Ordinary Least Square Estimation) : 잔차(Residual)를 2차 함수의 형태로 만들고, 이를 최소화하는 점에서 기울기와 절편 즉, 회귀계수(Coefficient)가 결정됨



# 참고 - 회귀분석의 기울기와 절편은 어떻게 찾을까?

## 회귀계수(Coefficient)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \quad \hat{\beta}_0 = \bar{y} - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \bar{x}$$

## 증명(Proof)

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \longrightarrow y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = e_i \quad (1)$$

$$(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = e_i^2 \quad (2)$$

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n e_i^2 \quad (3)$$

식 (3)의 우변을  $\hat{\beta}_1$ 에 대해 미분하면,

$$2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

식 (3)의 우변을  $\hat{\beta}_0$ 에 대해 미분하면,

$$2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow \hat{\beta}_0 = \bar{y} - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \bar{x}$$

잔차의 제곱합을 최소로 한다는 조건으로  
기울기에 해당하는  $\hat{\beta}_1$ 과 절편에 해당하는  
 $\hat{\beta}_0$ 을 구했으므로 회귀선(Regression  
line)을 그릴 수 있음