Chapter. 01

EDA & 회귀 분석

지도학습과 회귀 분석의 이해

FAST CAMPUS ONLINE

직장인을 위한 파이썬 데이터분석

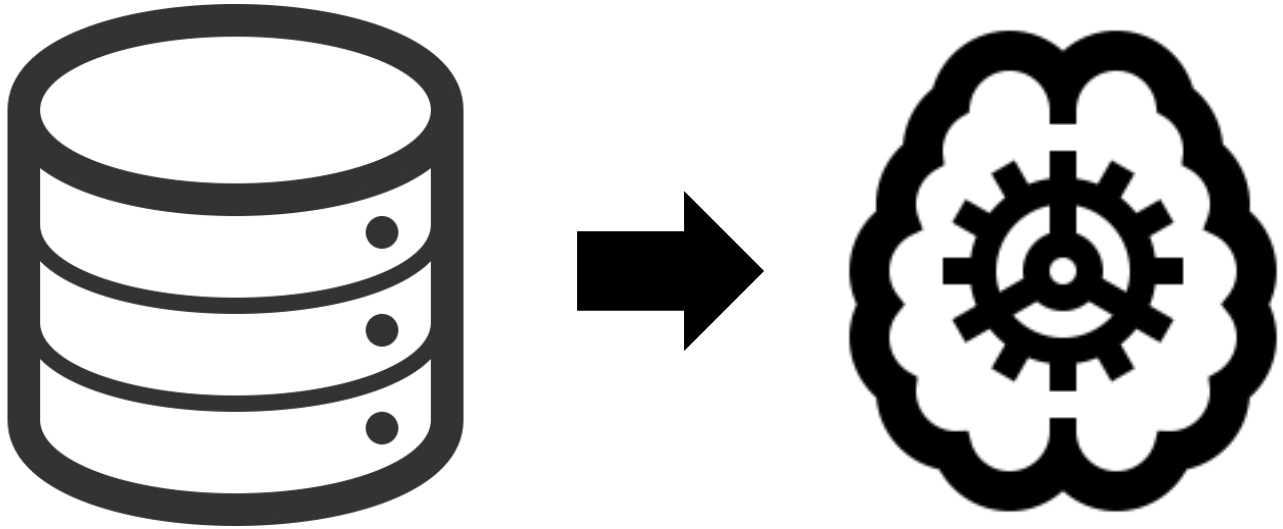
강사. 윤기태

Chapter. 01

지도 학습의 이해

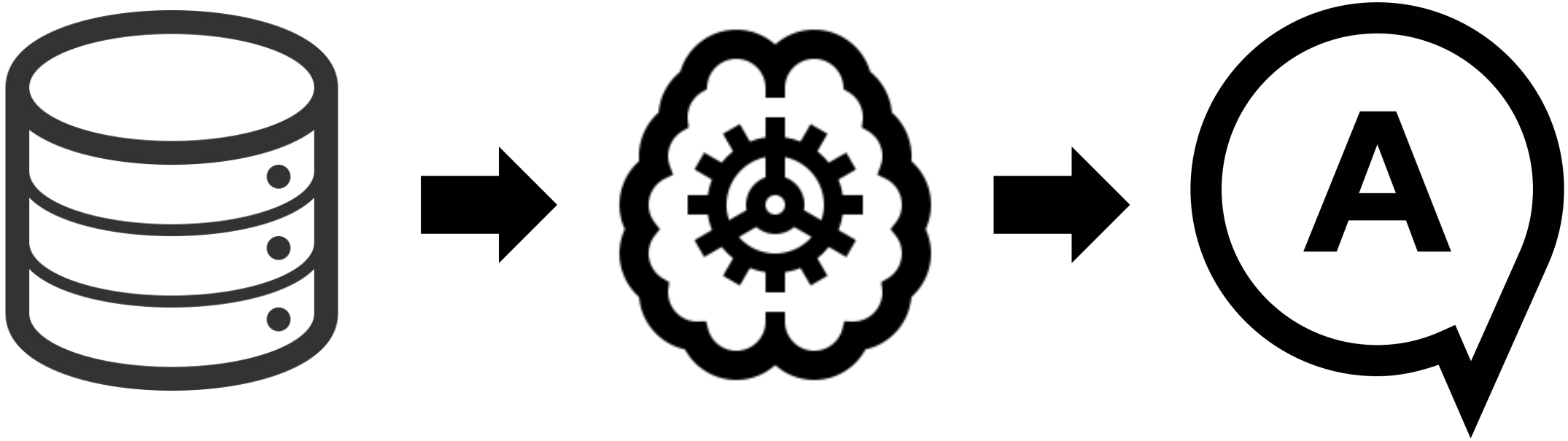
FAST CAMPUS
ONLINE
윤기태 강사.

I 지도 학습의 이해



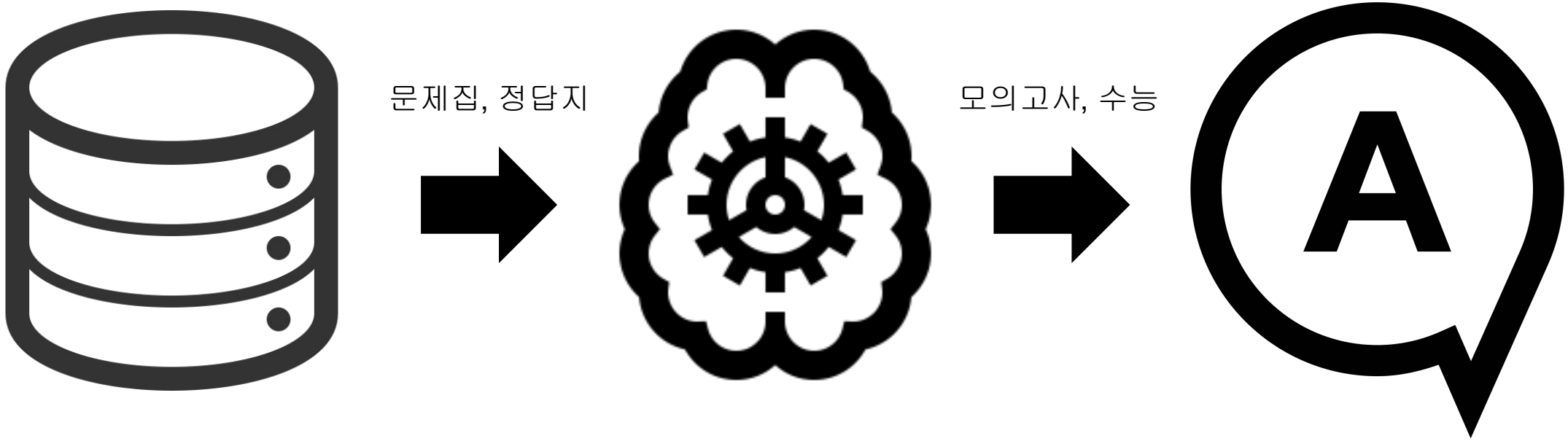
FAST CAMPUS
ONLINE
윤기태 강사.

I 지도 학습의 이해



FAST CAMPUS
ONLINE
윤기태 강사.

I 지도 학습의 이해



FAST CAMPUS
ONLINE
윤기태 강사.

I 지도 학습의 이해

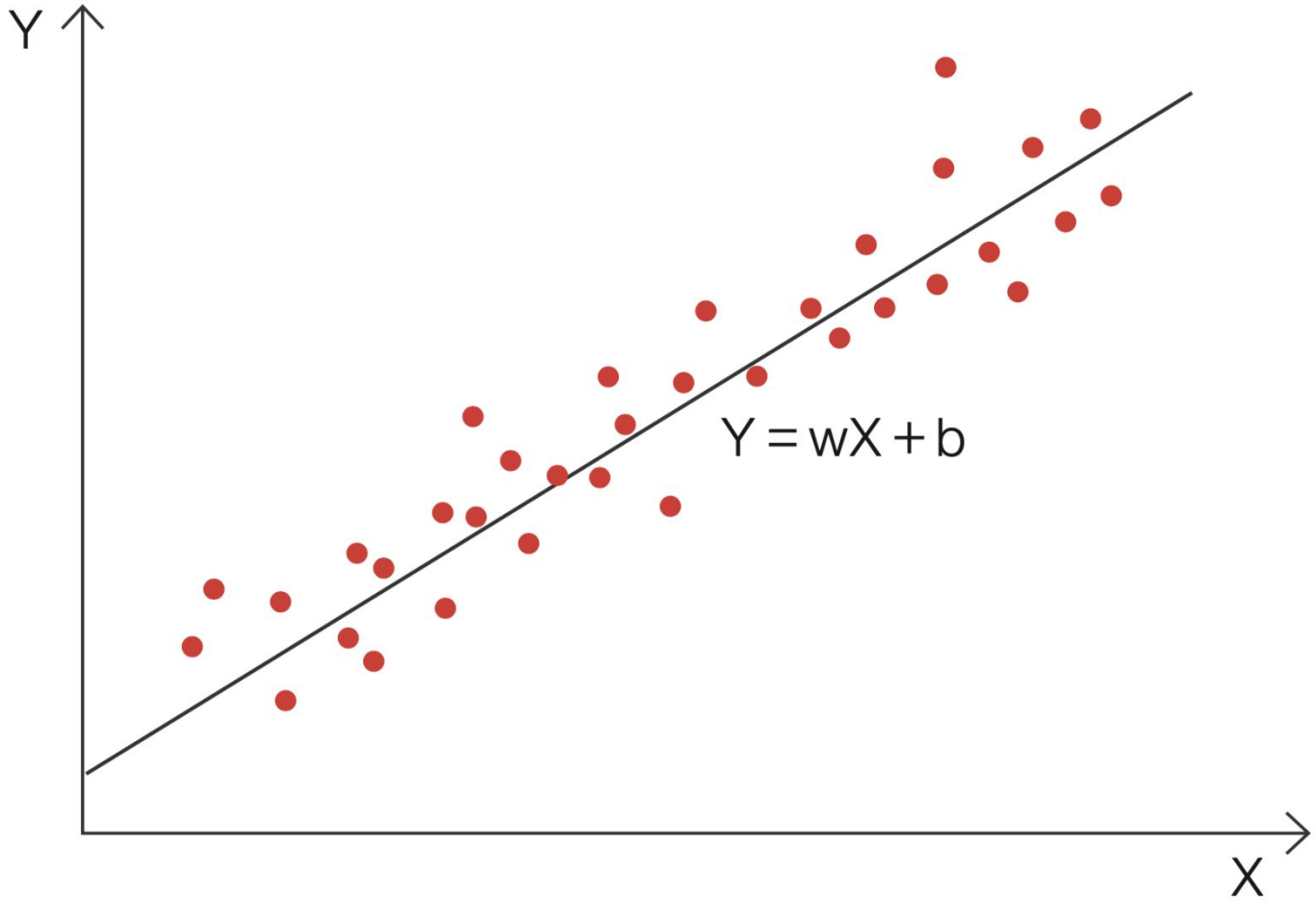
FAST CAMPUS
ONLINE
윤기태 강사.

Chapter. 01

회귀 분석의 이해

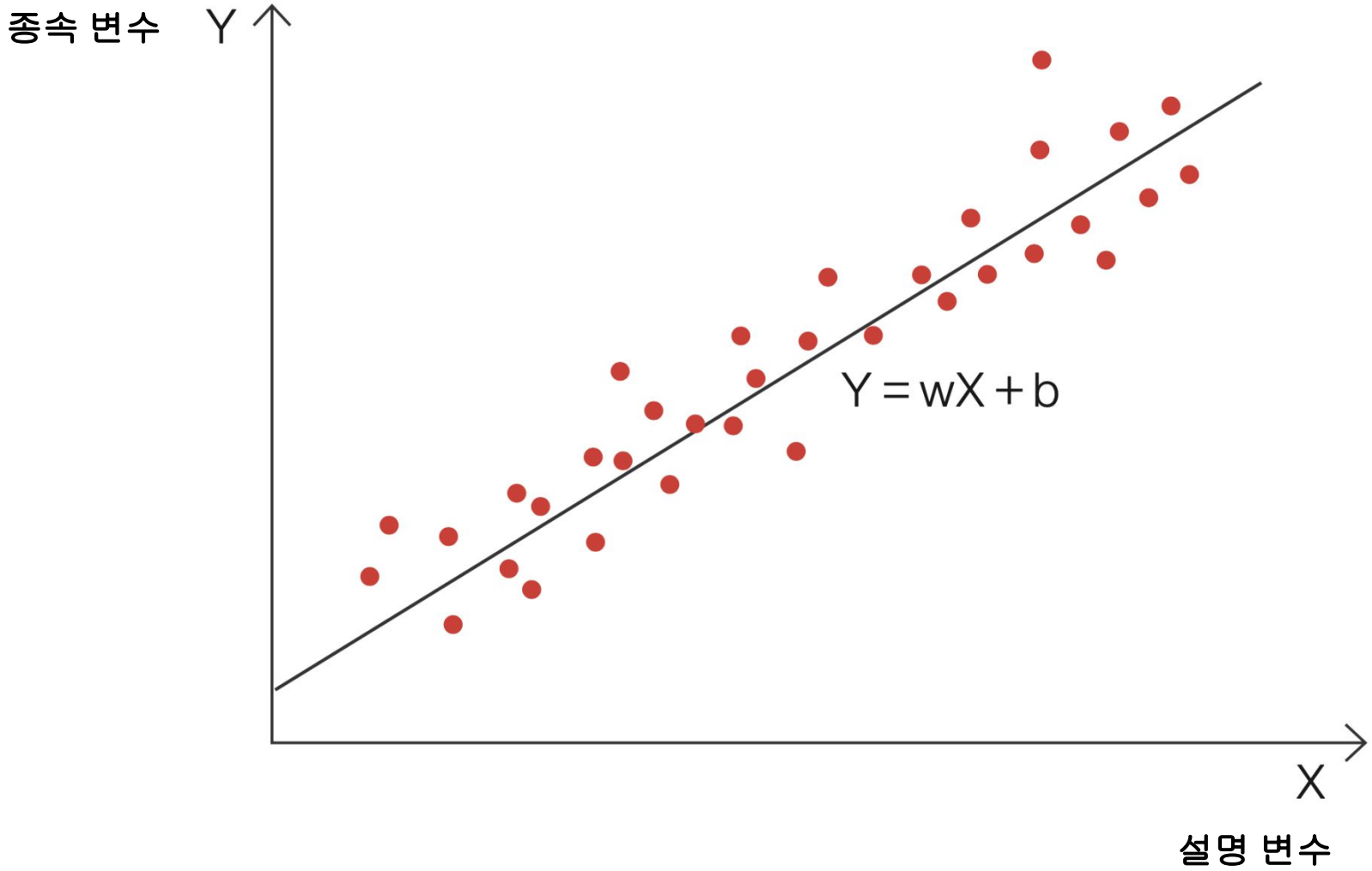
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 이해



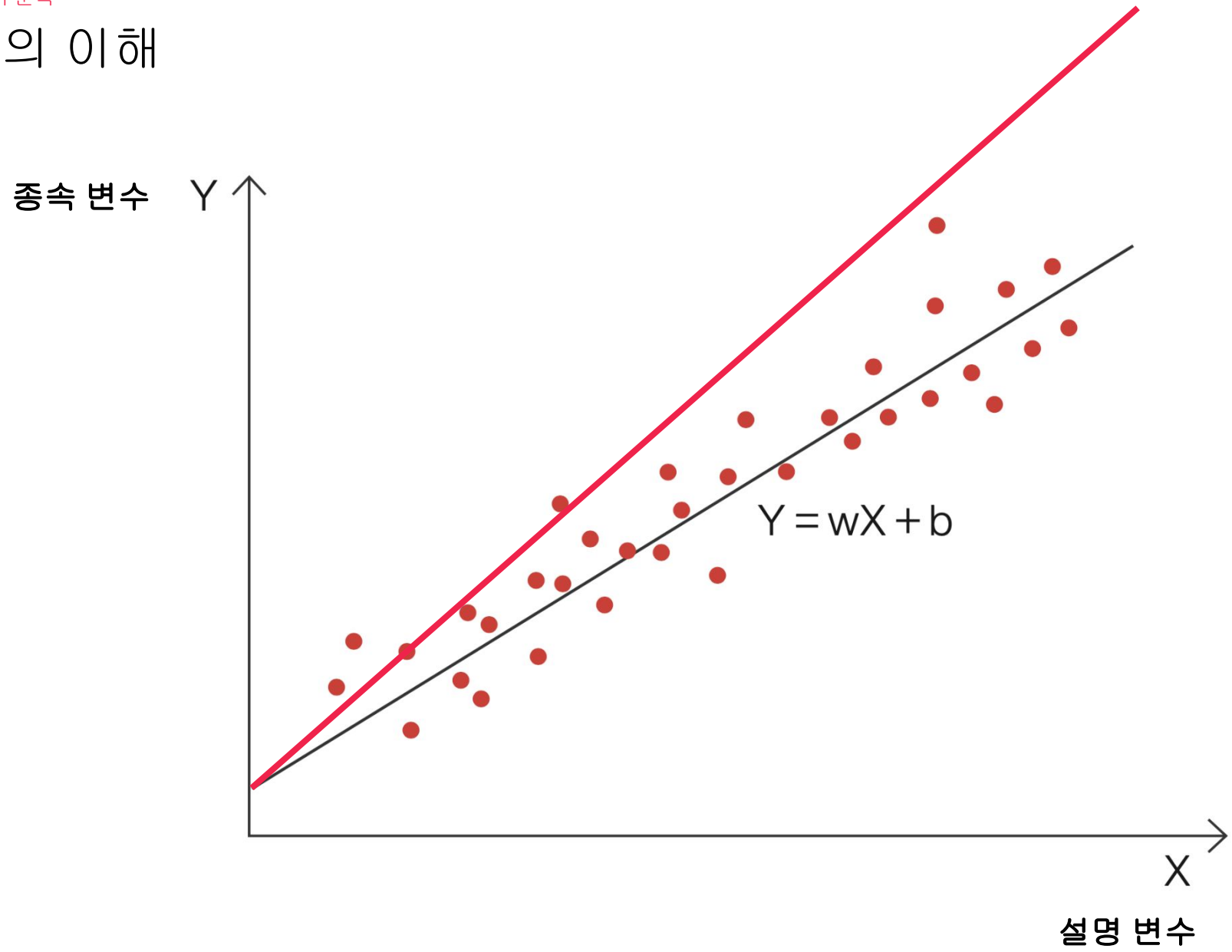
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 이해



FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 이해



FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 이해

설명 변수와 종속 변수간의 인과관계를 찾아내는 것

I 회귀 분석의 이해

Index	X1 (나이)	X2 (몸무게)	Y (키)
1	23	65.3	175.5
2	14	32.5	141.0
3	17	71.1	166.4
4	18	63.3	???

I 회귀 분석의 이해

$$Y = w1X1 + w2X2 + b$$

Index	X1 (나이)	X2 (몸무게)	Y (키)
1	23	65.3	175.5
2	14	32.5	141.0
3	17	71.1	166.4
4	18	63.3	???

I 회귀 분석의 이해

함수를 데이터에 맞추는 과정 **(모델 학습 과정)**

FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 이해

함수를 데이터에 맞추는 과정 **(모델 학습 과정)**

OLS(Ordinary Least Square) vs MLE(Maximum Likelihood Estimator)

Chapter. 01

모델 학습의 과정

I 회귀 분석 모델의 학습 과정

OLS (Ordinary Least Square)

제곱(Square)을 가장 작은(Least) 상태로 추정하는 것

I 회귀 분석 모델의 학습 과정

OLS (Ordinary Least Square)

제곱(Square)을 가장 작은(Least) 상태로 추정하는 것

=

I 회귀 분석 모델의 학습 과정

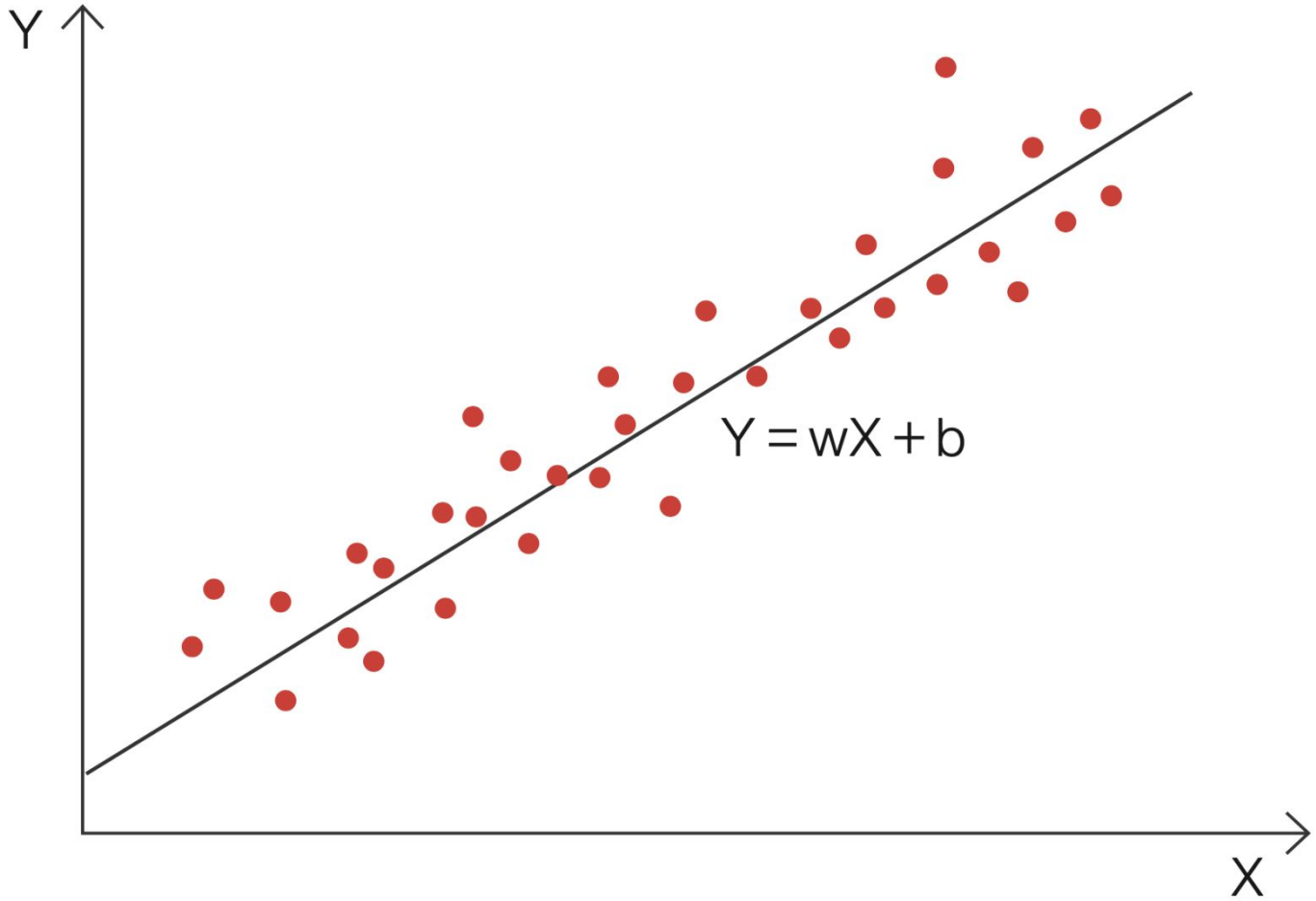
OLS (Ordinary Least Square)

제곱(Square)을 가장 작은(Least) 상태로 추정하는 것

=

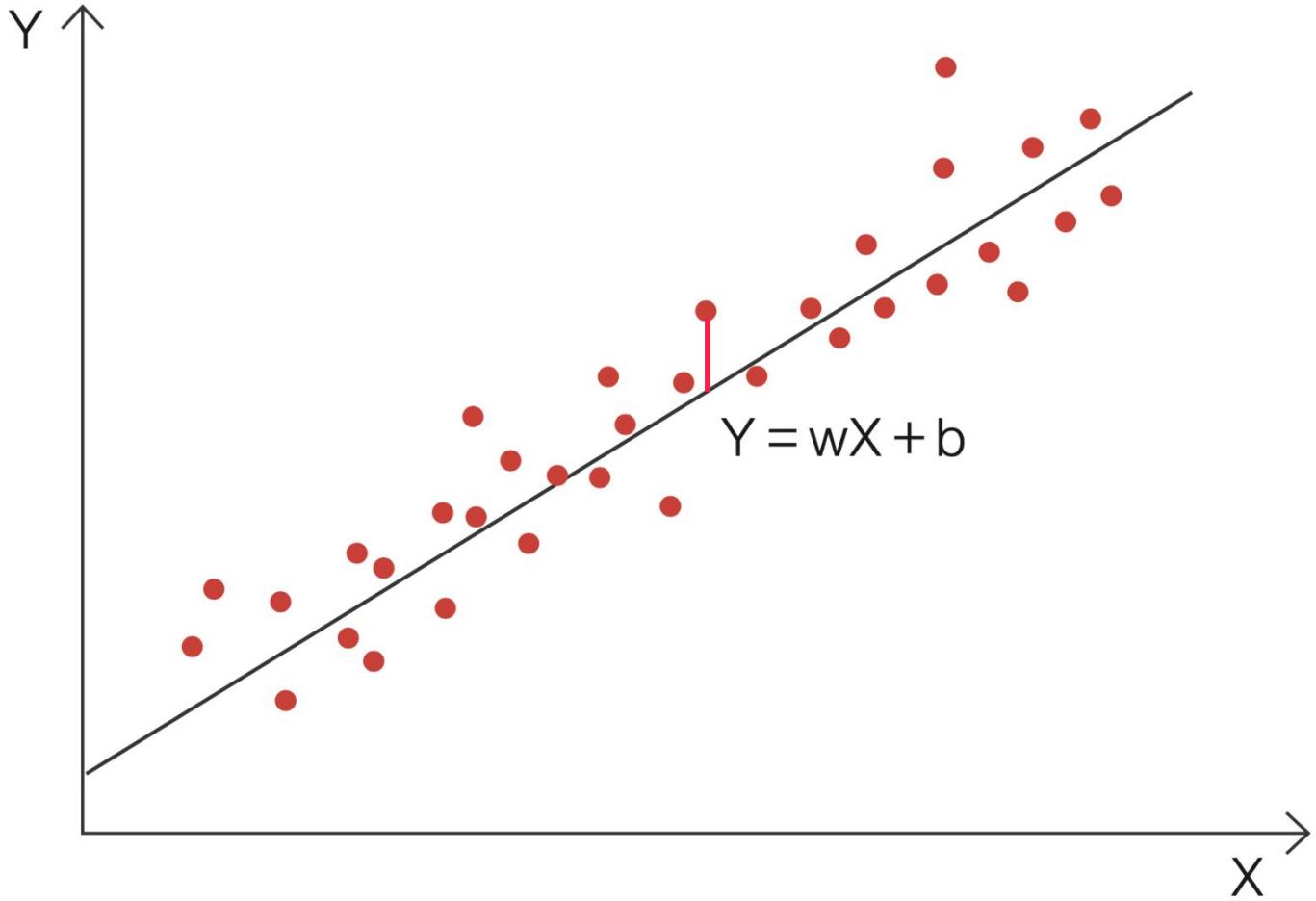
오차들의 제곱을 최소화 하는 것

I 회귀 분석 모델의 학습 과정



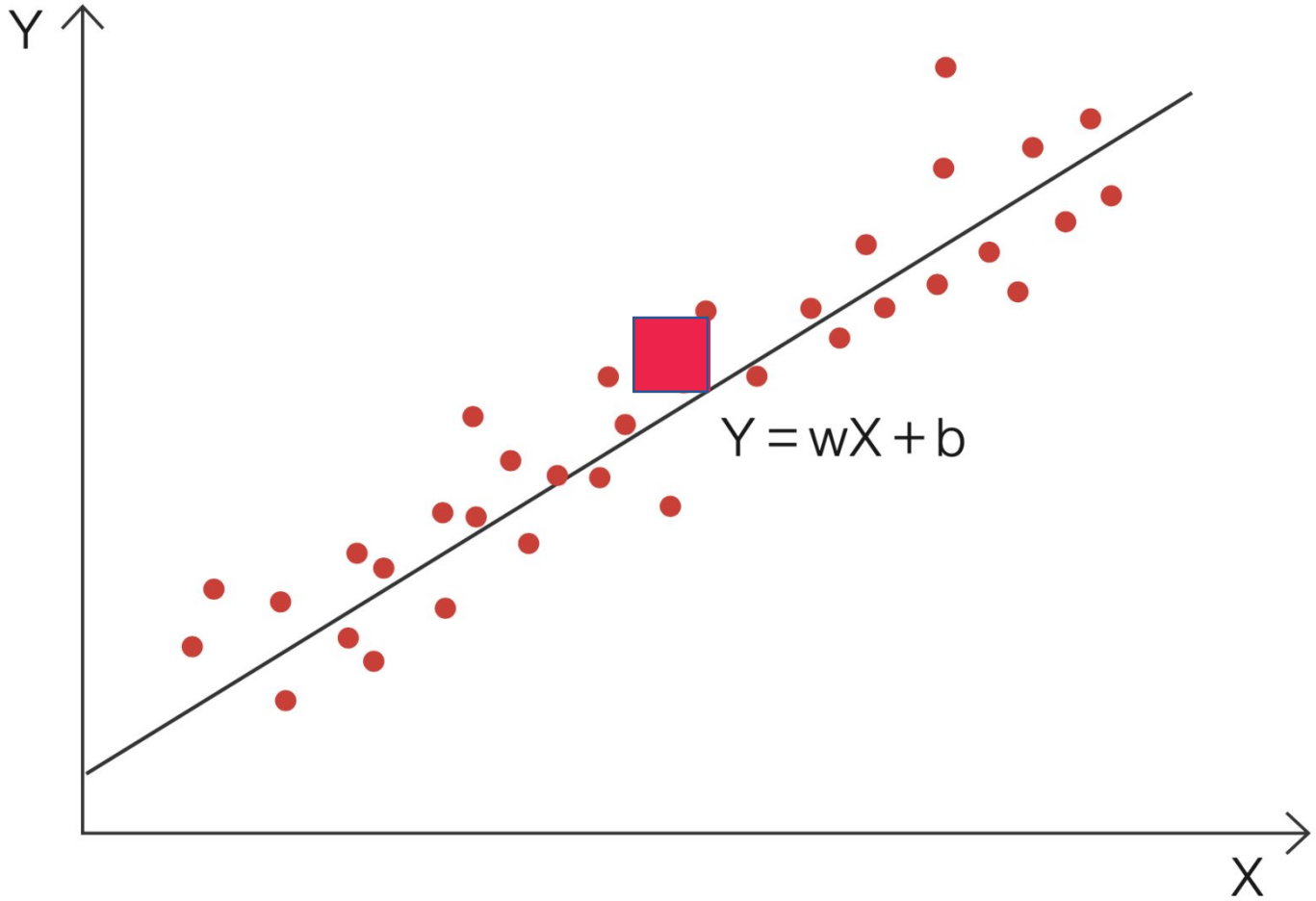
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델의 학습 과정



FAST CAMPUS
ONLINE
윤기태 강사.

회귀 분석 모델의 학습 과정



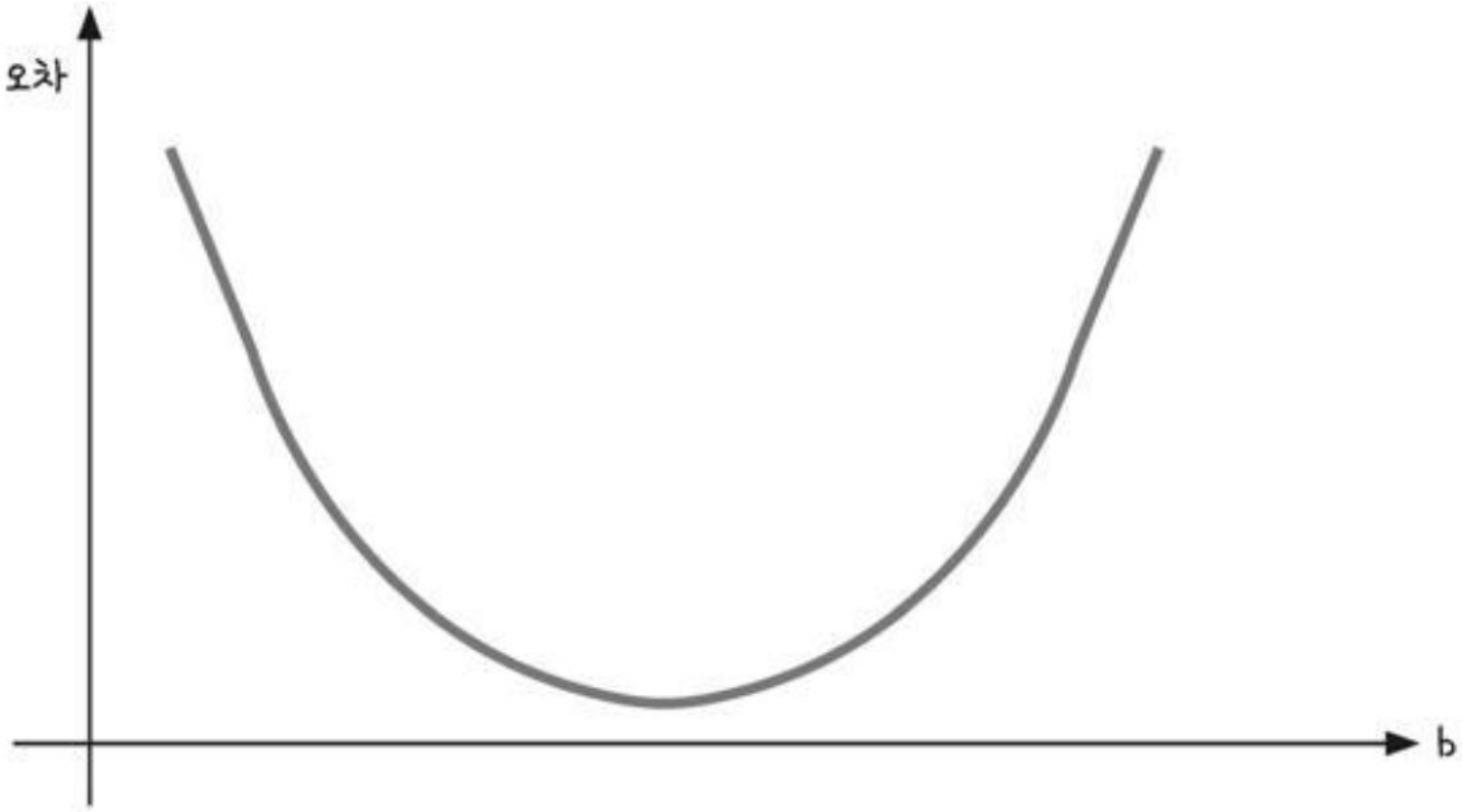
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델의 학습 과정

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

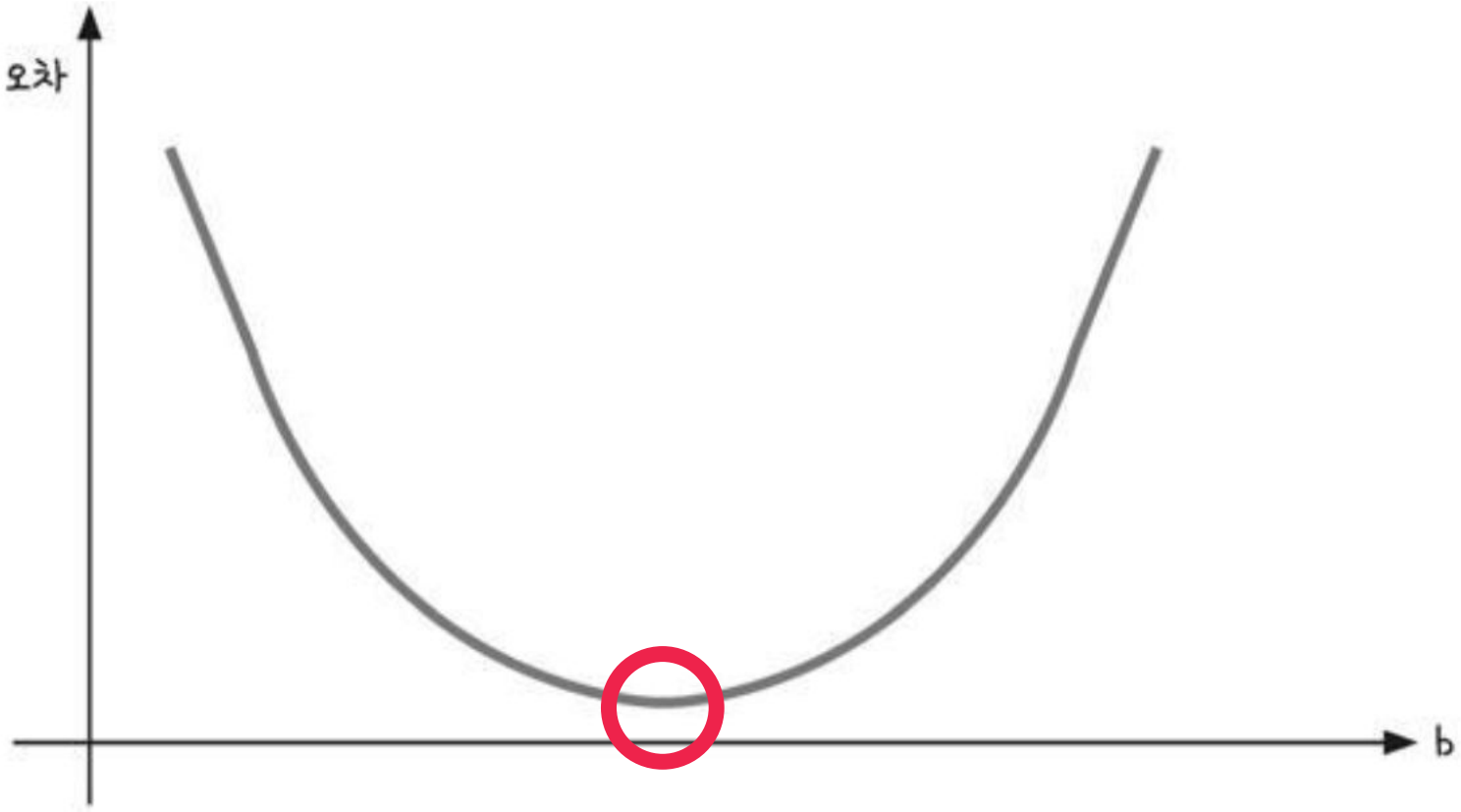
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델의 학습 과정



FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델의 학습 과정



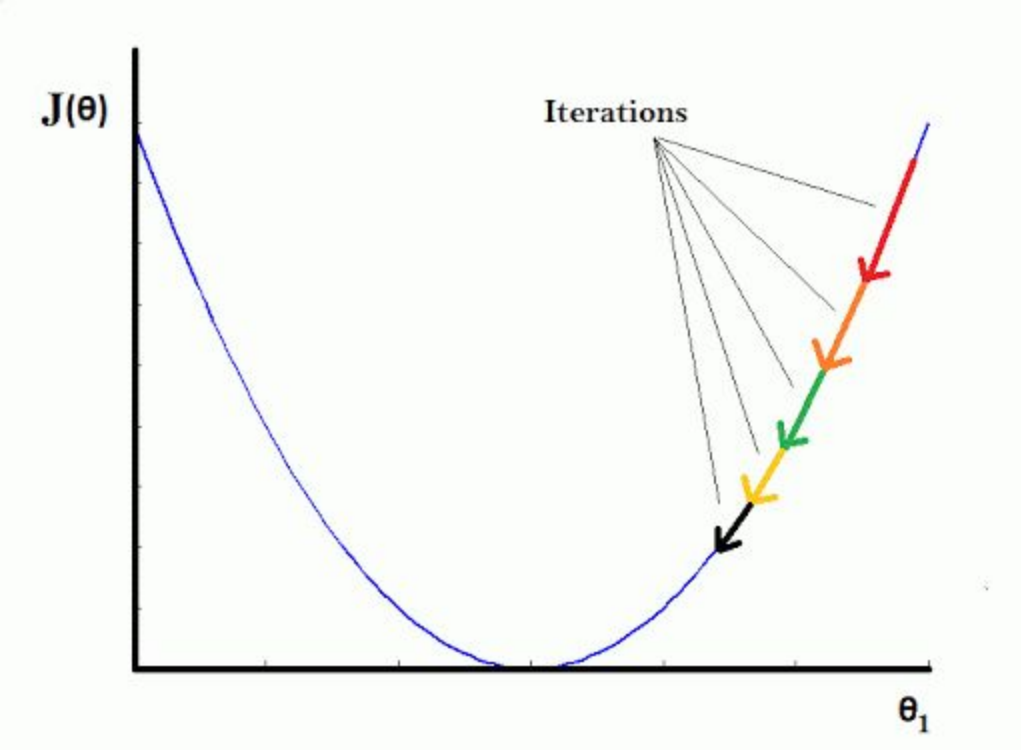
FAST CAMPUS
ONLINE
윤기태 강사.

I 참고 : 그래디언트 디센트의 이해

2. 그래디언트 디센트(Gradient Descent)

학습을 반복할때마다 줄어드는 오차를
반영하여 m과 b를 조금씩 근사시키는 방법

I 참고 : 그래디언트 디센트의 이해



Chapter. 01

모델 평가 방법

FAST CAMPUS
ONLINE
윤기태 강사.

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	**R-squared (uncentered):**	0.137
Model:	OLS	**Adj. R-squared (uncentered):**	0.111
Method:	Least Squares	**F-statistic:**	5.203
Date:	Mon, 09 Mar 2020	**Prob (F-statistic):**	3.95e-08
Time:	14:50:03	**Log-Likelihood:**	-1838.2
No. Observations:	404	**AIC:**	3700.
Df Residuals:	392	**BIC:**	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	**Durbin-Watson:**	0.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	509.230
Skew:	1.516	**Prob(JB):**	2.64e-111
Kurtosis:	7.589	**Cond. No.**	9.50

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	R-squared (uncentered):	0.137
Model:	OLS	Adj. R-squared (uncentered):	0.111
Method:	Least Squares	F-statistic:	5.203
Date:	Mon, 09 Mar 2020	Prob (F-statistic):	3.95e-08
Time:	14:50:03	Log-Likelihood:	-1838.2
No. Observations:	404	AIC:	3700.
Df Residuals:	392	BIC:	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	Durbin-Watson:	0.097
Prob(Omnibus):	0.000	Jarque-Bera (JB):	509.230
Skew:	1.516	Prob(JB):	2.64e-111
Kurtosis:	7.589	Cond. No.	9.50

I 모델 평가 방법

R-squared (결정 계수)

I 모델 평가 방법

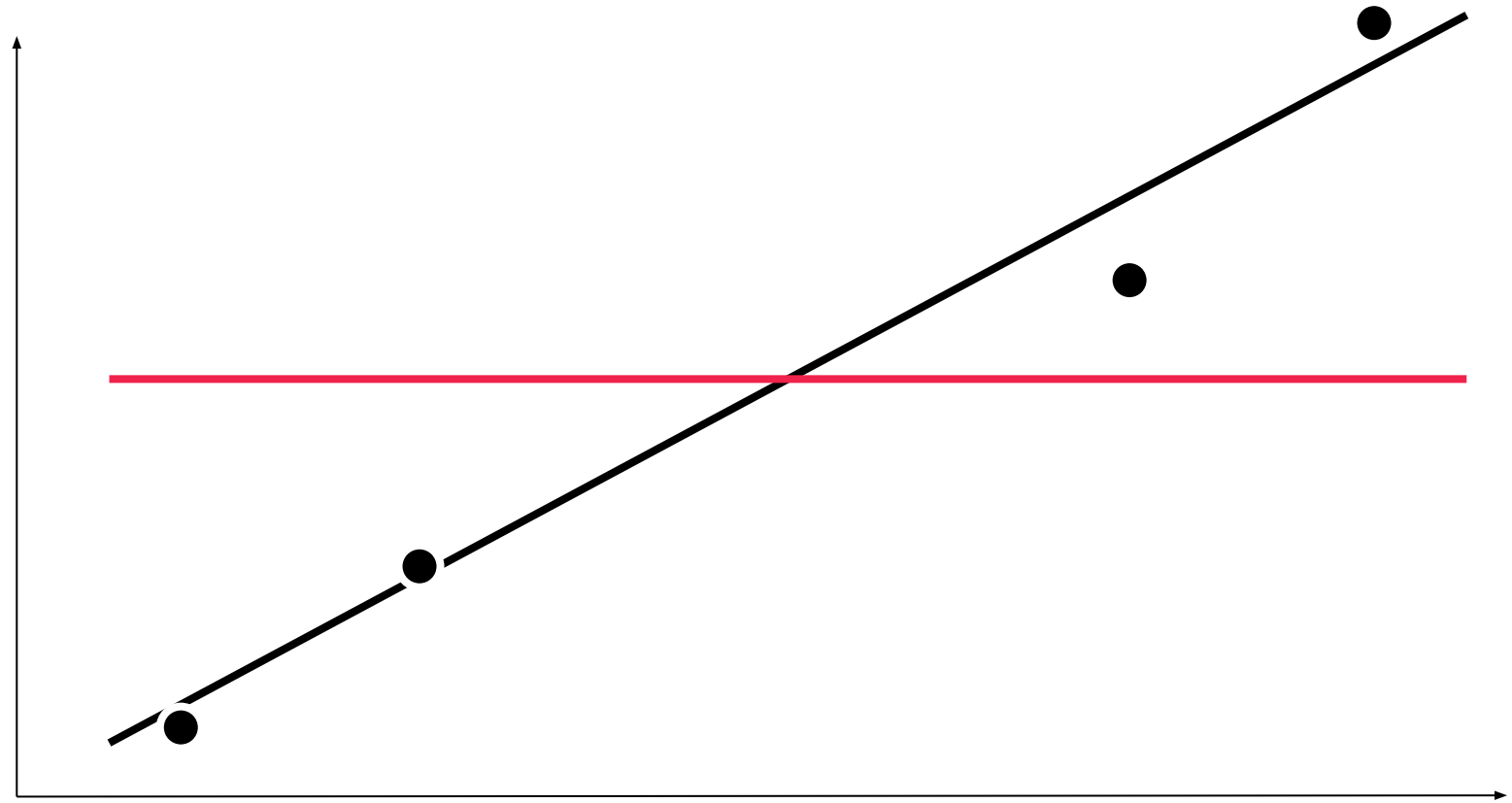
R-squared (결정 계수)

=

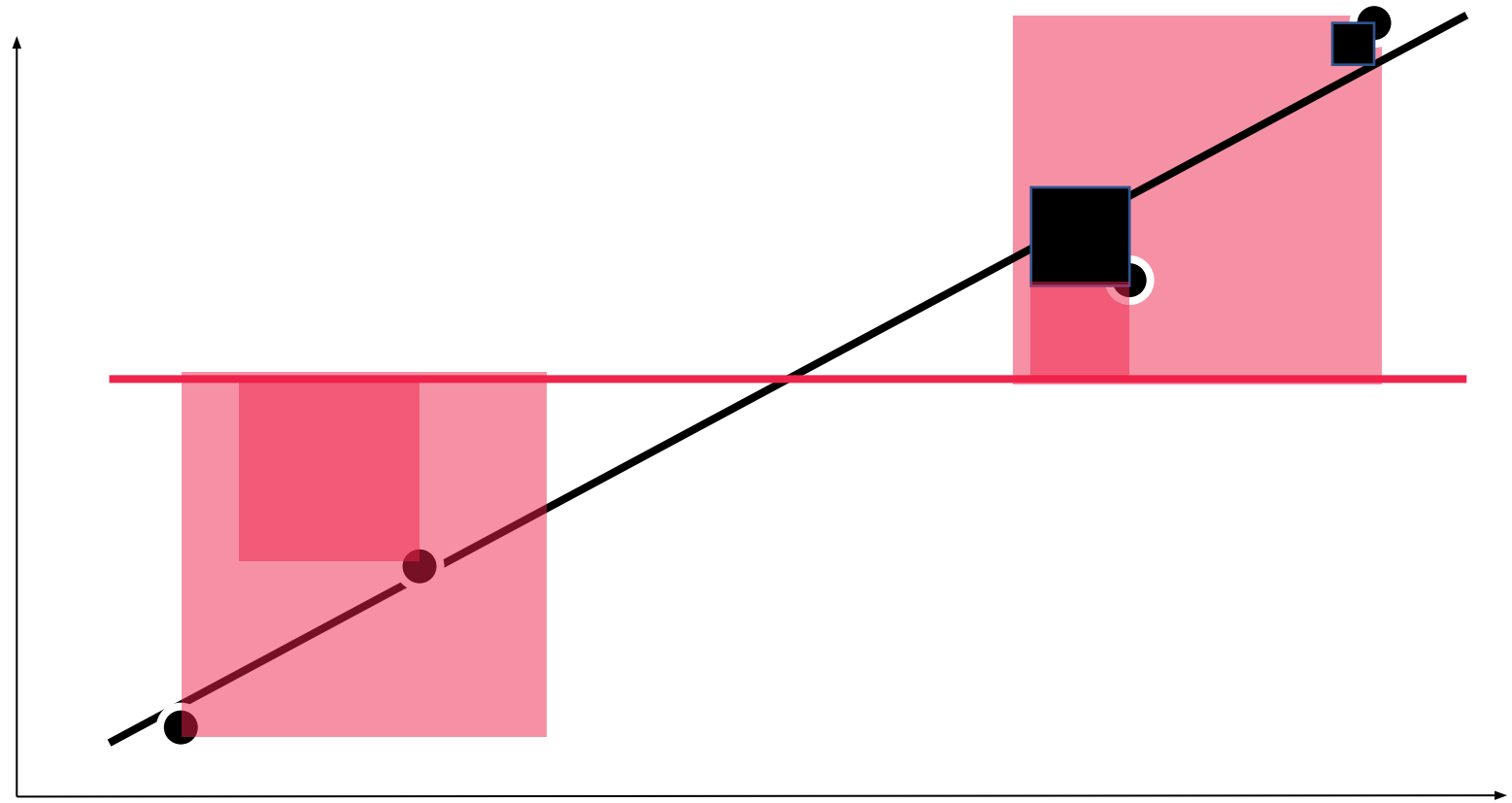
$$R^2 = 1 - \frac{MSE\ of\ regression\ line}{MSE\ of\ the\ average\ of\ the\ data}$$

“데이터의 점들을 얼마나 잘 설명하고 있는가”

I 모델 평가 방법



I 모델 평가 방법

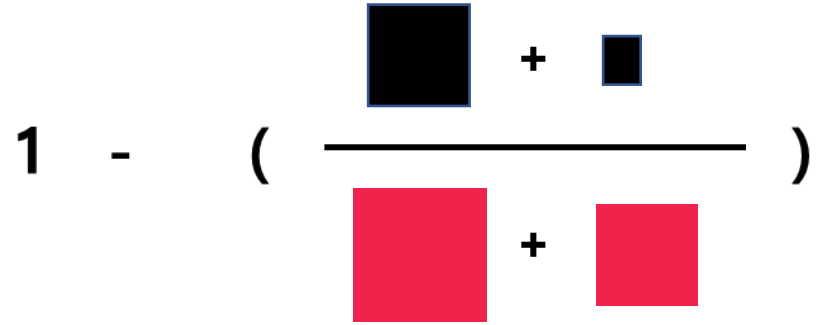


I 모델 평가 방법

R-squared (결정 계수)

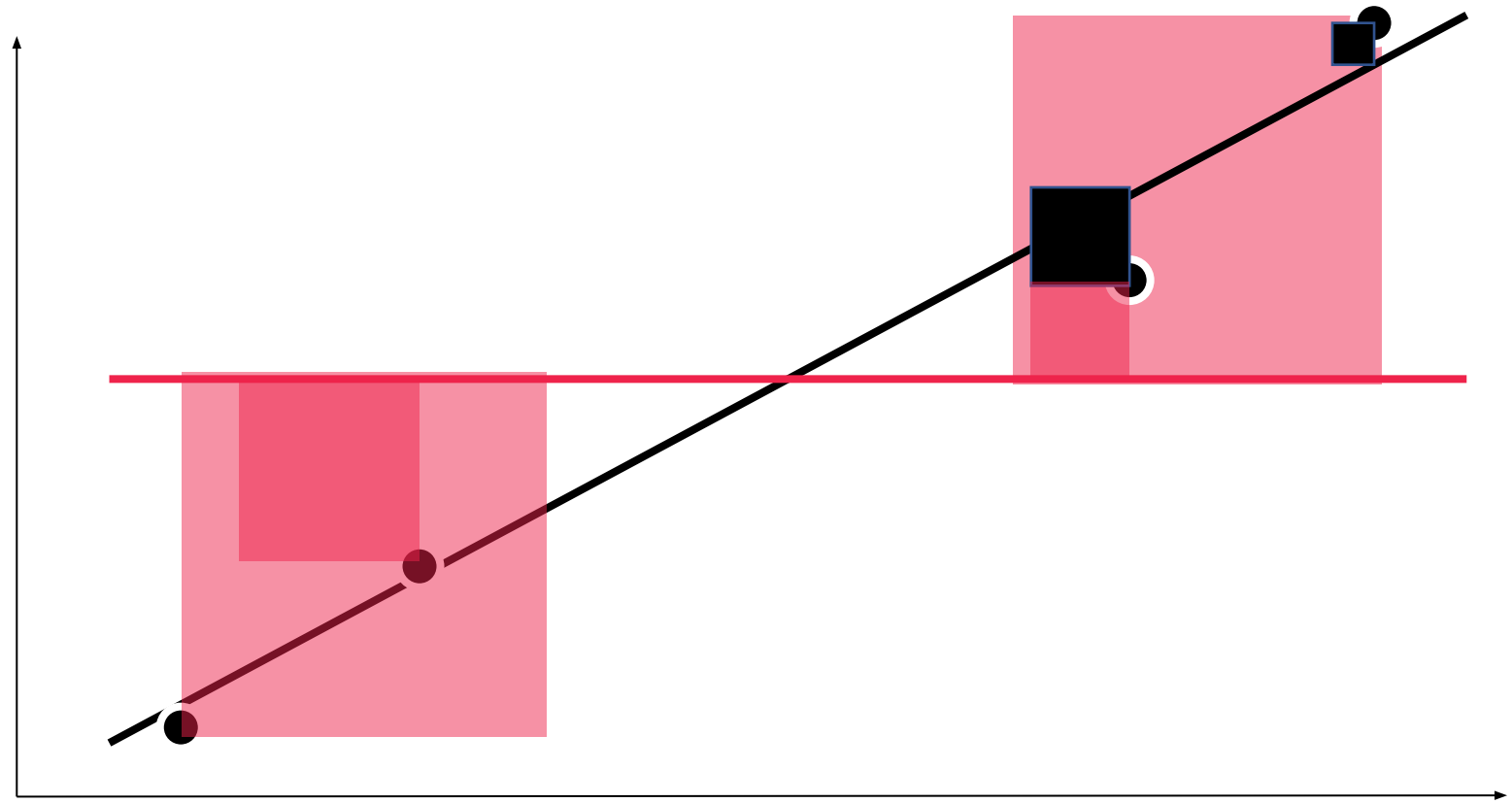
=

$$R^2 = 1 - \frac{MSE\ of\ regression\ line}{MSE\ of\ the\ average\ of\ the\ data}$$



FAST CAMPUS
ONLINE
윤기태 강사.

I 모델 평가 방법



I 모델 평가 방법

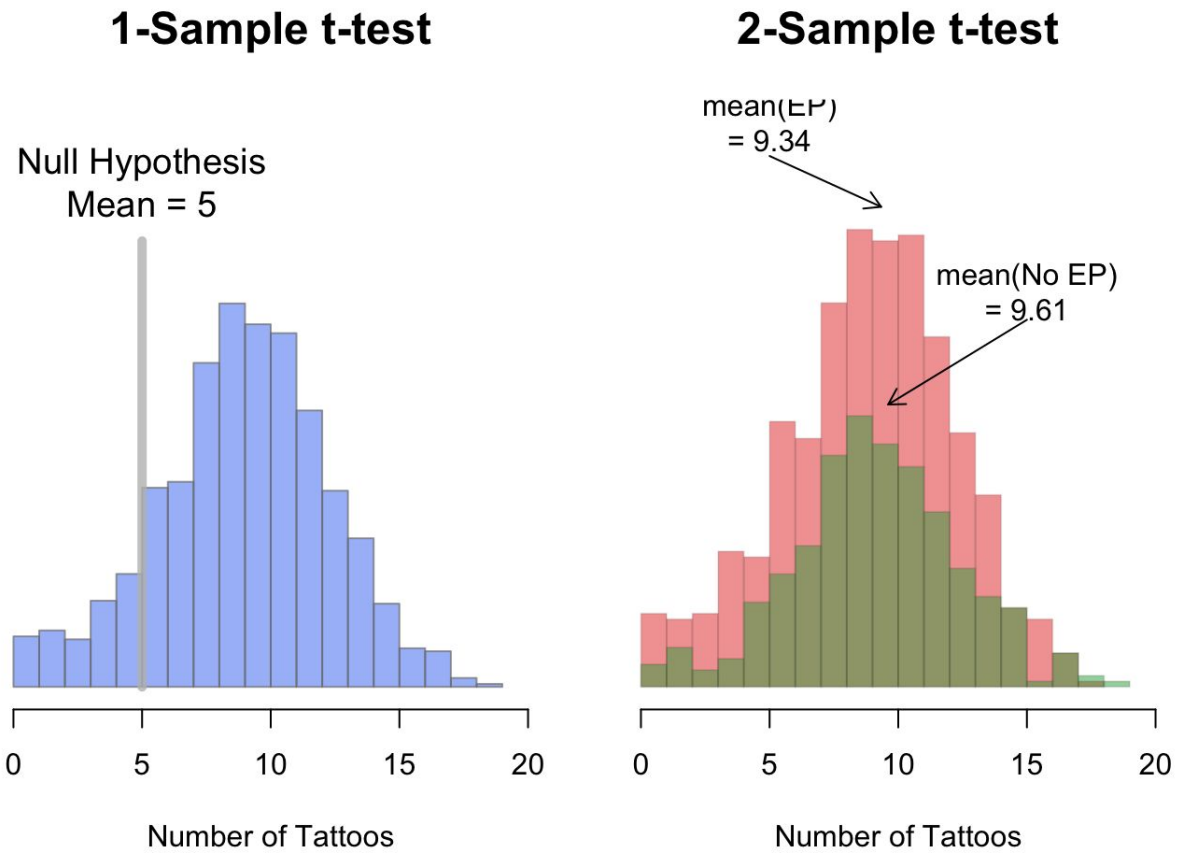
검정 방법과 유의성

FAST CAMPUS
ONLINE
윤기태 강사.

I 모델 평가 방법

검정 방법과 유의성

예시) T-test



FAST CAMPUS
ONLINE
윤기태 강사.

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	**R-squared (uncentered):**	0.137
Model:	OLS	**Adj. R-squared (uncentered):**	0.111
Method:	Least Squares	**F-statistic:**	5.203
Date:	Mon, 09 Mar 2020	**Prob (F-statistic):**	3.95e-08
Time:	14:50:03	**Log-Likelihood:**	-1838.2
No. Observations:	404	**AIC:**	3700.
Df Residuals:	392	**BIC:**	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	**Durbin-Watson:**	0.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	509.230
Skew:	1.516	**Prob(JB):**	2.64e-111
Kurtosis:	7.589	**Cond. No.**	9.50

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	**R-squared (uncentered):**	0.137
Model:	OLS	**Adj. R-squared (uncentered):**	0.111
Method:	Least Squares	**F-statistic:**	5.203
Date:	Mon, 09 Mar 2020	**Prob (F-statistic):**	3.95e-08
Time:	14:50:03	**Log-Likelihood:**	-1838.2
No. Observations:	404	**AIC:**	3700.
Df Residuals:	392	**BIC:**	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	**Durbin-Watson:**	0.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	509.230
Skew:	1.516	**Prob(JB):**	2.64e-111
Kurtosis:	7.589	**Cond. No.**	9.50

P-value (Prob, P)

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	**R-squared (uncentered):**	0.137
Model:	OLS	**Adj. R-squared (uncentered):**	0.111
Method:	Least Squares	**F-statistic:**	5.203
Date:	Mon, 09 Mar 2020	**Prob (F-statistic):**	3.95e-08
Time:	14:50:03	**Log-Likelihood:**	-1838.2
No. Observations:	404	**AIC:**	3700.
Df Residuals:	392	**BIC:**	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	**Durbin-Watson:**	0.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	509.230
Skew:	1.516	**Prob(JB):**	2.64e-111
Kurtosis:	7.589	**Cond. No.**	9.50

????

I 모델 평가 방법

OLS Regression Results

Dep. Variable:	CMEDV	**R-squared (uncentered):**	0.137
Model:	OLS	**Adj. R-squared (uncentered):**	0.111
Method:	Least Squares	**F-statistic:**	5.203
Date:	Mon, 09 Mar 2020	**Prob (F-statistic):**	3.95e-08
Time:	14:50:03	**Log-Likelihood:**	-1838.2
No. Observations:	404	**AIC:**	3700.
Df Residuals:	392	**BIC:**	3748.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
CRIM	-0.5092	1.461	-0.349	0.728	-3.381	2.363
ZN	1.3977	1.726	0.810	0.418	-1.995	4.790
INDUS	0.6053	2.286	0.265	0.791	-3.888	5.099
NOX	-3.1767	2.503	-1.269	0.205	-8.098	1.744
RM	2.6496	1.590	1.666	0.096	-0.477	5.776
AGE	0.2291	1.998	0.115	0.909	-3.699	4.158
DIS	-4.9476	2.397	-2.064	0.040	-9.659	-0.236
RAD	2.6207	3.148	0.833	0.406	-3.568	8.809
TAX	-3.5000	3.473	-1.008	0.314	-10.327	3.327
PTRATIO	-1.5754	1.516	-1.039	0.299	-4.556	1.405
B	1.8915	1.427	1.326	0.186	-0.913	4.696
LSTAT	-4.6766	1.974	-2.369	0.018	-8.557	-0.796

Omnibus:	138.790	**Durbin-Watson:**	0.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	509.230
Skew:	1.516	**Prob(JB):**	2.64e-111
Kurtosis:	7.589	**Cond. No.**	9.50

w가 0인지 아닌지에 대한 검정

I 모델 평가 방법

다중 공선성

"변수간의 강한 상관관계가 발생한 경우"