

17기 정규세션

ToBig's 16기강의자
이예림

Regression Analysis

회귀분석

Contents

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가 지표

Intro

1. 머신러닝 알고리즘

지도학습 (Supervised Learning)	비지도학습 (Unsupervised Learning)	강화학습 (Reinforcement Learning)
<ul style="list-style-type: none">• 입력과 결과값(Label) 이용한 학습• 회귀(Regression)• 분류(Classification)	<ul style="list-style-type: none">• 입력만을 이용한 학습• 군집화(Clustering)	<ul style="list-style-type: none">• Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습
Ex. 선형 회귀, 로지스틱 회귀 , KNN, SVM, Decision Tree	Ex. K-Means Clustering	

Intro

2. 인과관계 VS 상관관계



인과관계(Causality)

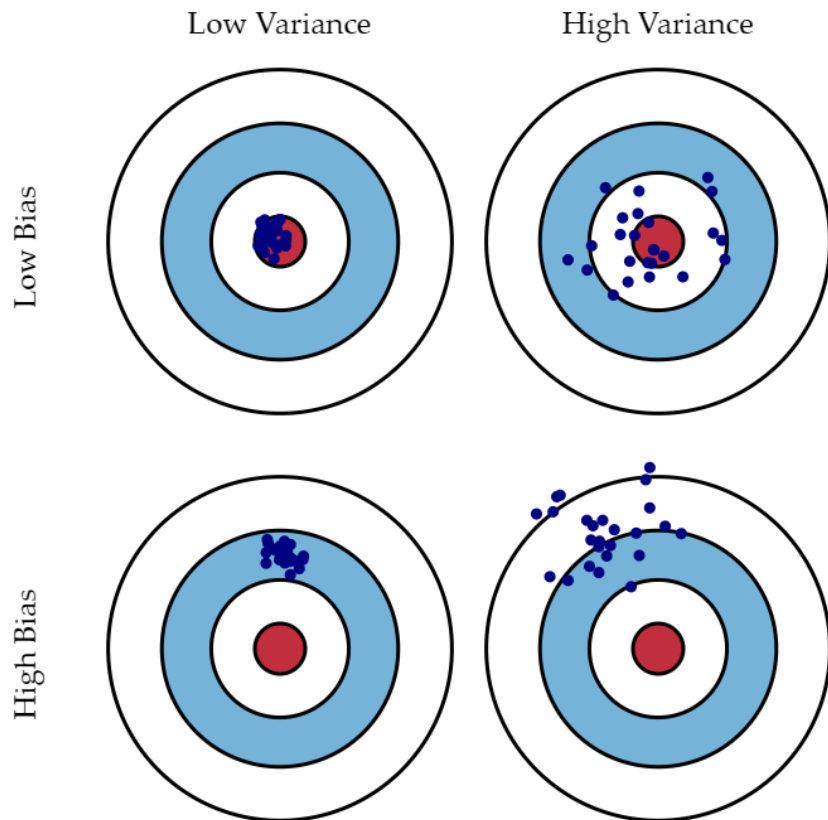
- 어떤 사실과 다른 사실 사이의 원인과 결과 관계

상관관계(Association, Correlation)

- 두 변량 중 한쪽이 증가함에 따라, 다른 한 쪽이 증가 또는 감소하는 관계
- 상관관계가 존재할 때, 필연적으로 인과관계가 존재하는 것은 아님

Intro

3. 편향(Bias) VS 분산(Variance)



Bias

- 데이터 내 모든 정보를 고려하지 않기에, 알고리즘이 지속적으로 잘못된 내용을 학습하는 경향성
- **Underfitting과 관련**

Variance

- Highly flexible model에 데이터를 fit함으로써, 실제 현상과 관계 없는 random한 것들까지 학습하는 알고리즘의 경향성
- **Overfitting과 관련**

★ Bias-Variance Trade-off

Contents

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

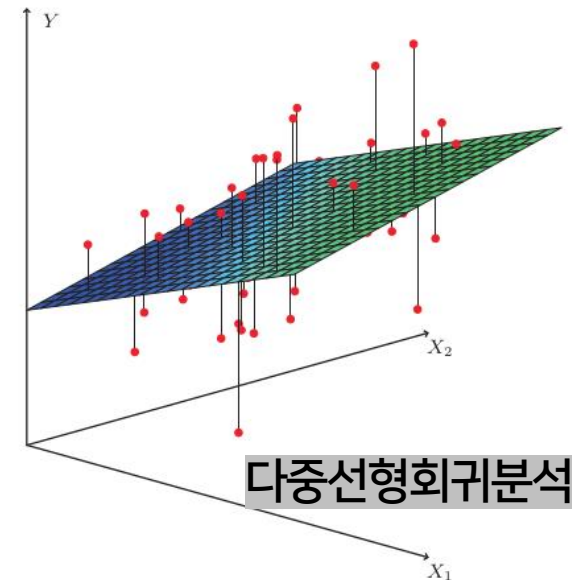
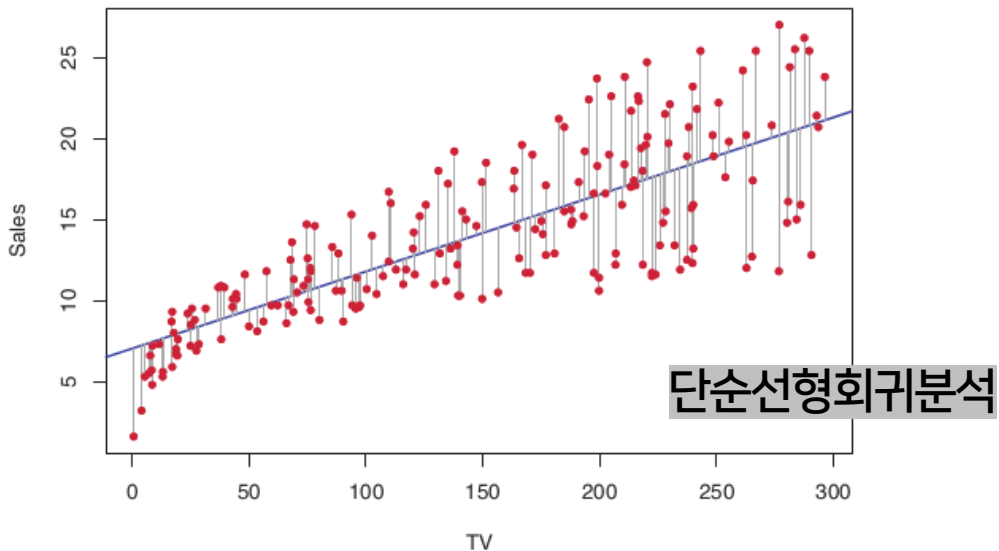
Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가 지표

Unit 01 | 선형 회귀분석

선형 회귀분석 (Linear Regression)

- **회귀분석** : 설명변수(X)에 대응하는 반응변수(Y)와 가장 비슷한 값(\hat{Y})을 출력하는 함수를 찾는 과정
- **선형 회귀분석** : 반응변수와 한 개 이상의 설명변수와의 선형 상관관계를 모델링하는 회귀분석 기법
 - Ex. 시험 공부 시간(X)에 따른 시험 성적(Y)



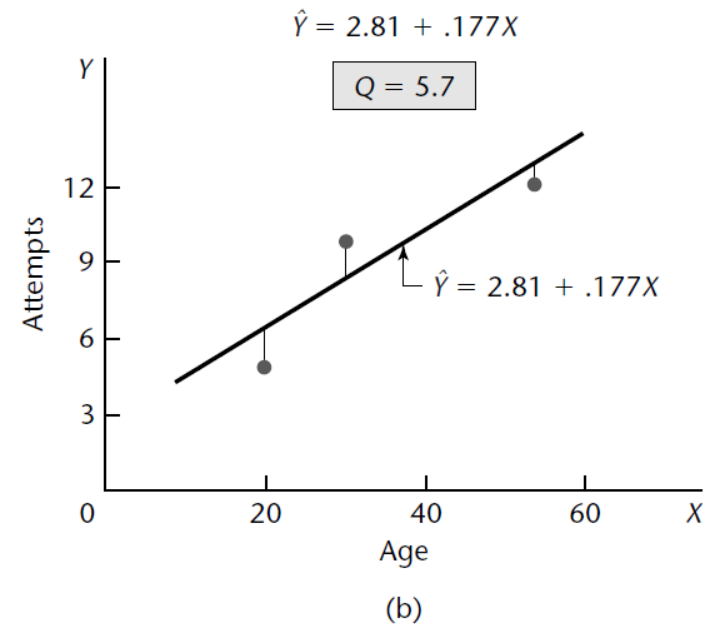
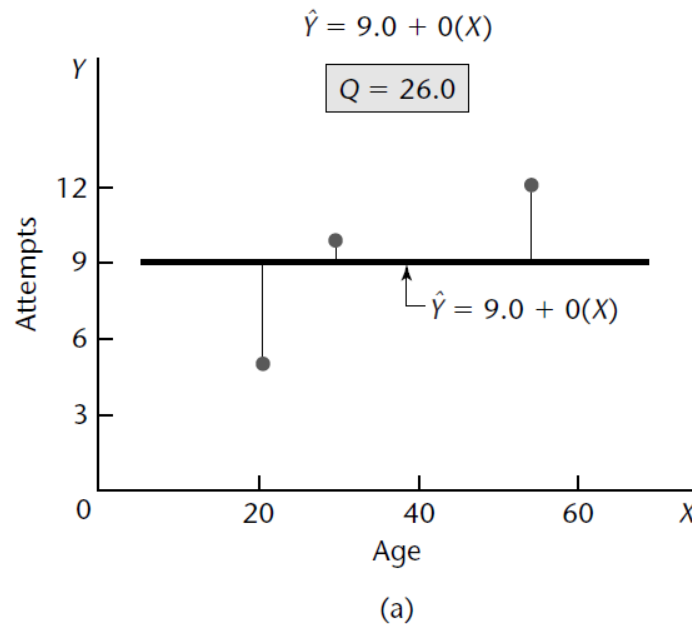
Unit 01 | 선형 회귀분석

Formulation : $Y = \beta_0 + \beta_1 X + \epsilon$ (※ 선형 상관관계 가정)

- β_0, β_1 : 회귀계수(regression coefficients), unknown but random
- $\widehat{\beta}_0, \widehat{\beta}_1$: 예측된 회귀계수 \rightarrow 예측값 : $\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$
- Loss = 실제값(Y) - 예측값(\hat{Y}) \rightarrow Loss를 최소화하는 방법이 최소제곱법(LSE)

Unit 01 | 선형 회귀분석

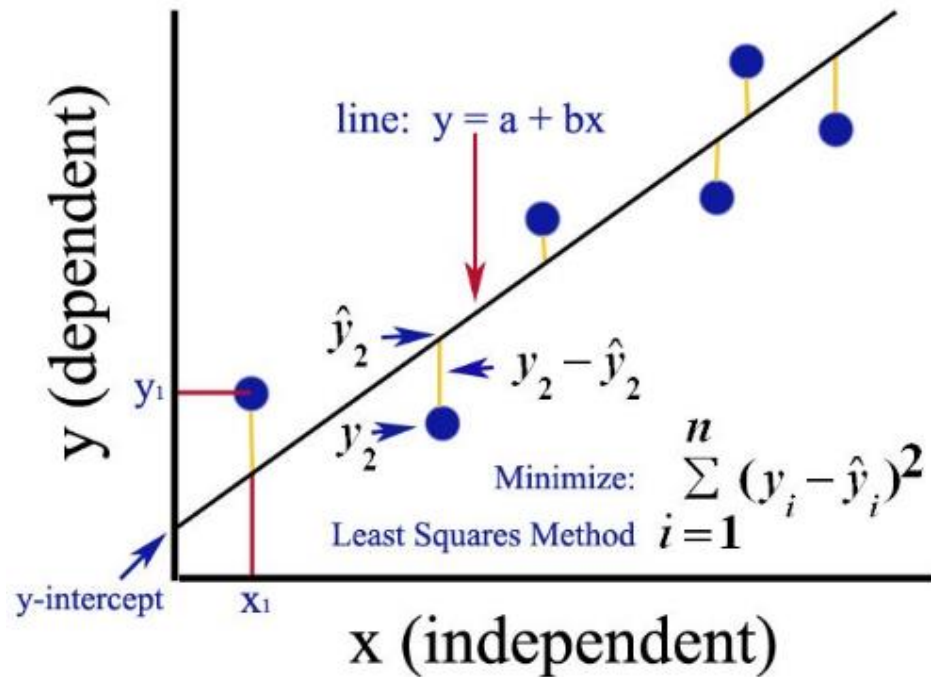
Q. 주어진 데이터에 대해 더 optimal한 line은?



직관적으로, (b)가 더욱 optimal한 line!
∴ (b)가 데이터와 회귀직선 간 거리가 더 가까움

Unit 01 | 선형 회귀분석

최소제곱법 (LSE : Least Squares Estimation)



예측값(\hat{Y})과 실제값(Y) 간 차이(=잔차)의 제곱합을 최소화하는 알고리즘 → 최적의 회귀계수(모수) 추정

Loss Function : $L = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$
→ L 이 작을수록 좋은 회귀모델!

Unit 01 | 선형 회귀분석

최소제곱법 (LSE : Least Squares Estimation) 단순선형회귀분석

$$Q = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

정규방정식
(Normal Equation)목적함수(Q)를 최소화하기
위한 편미분

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= -2 \sum (Y_i - \beta_0 - \beta_1 X_i) \\ \frac{\partial Q}{\partial \beta_1} &= -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)\end{aligned}$$

정규방정식을 풀면, 다음과 같은 결과를 얻게 된다.

최소제곱 추정치(Least Squares Estimator)

$$\begin{aligned}b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ b_0 &= \frac{1}{n} \left(\sum Y_i - b_1 \sum X_i \right) = \bar{Y} - b_1 \bar{X}\end{aligned}$$

Unit 01 | 선형 회귀분석

최소제곱법 (LSE : Least Squares Estimation) 다중선형회귀분석

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

(6.18a)

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

(6.18b)

$$\mathbf{X}_{n \times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

(6.18c)

$$\boldsymbol{\beta}_{p \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

(6.18d)

$$\boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

행렬표현

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

목적함수(Q)를 최소화하기
위한 **편미분**

정규방정식

$$\frac{\partial}{\partial \boldsymbol{\beta}} (Q) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

정규방정식을 풀면,

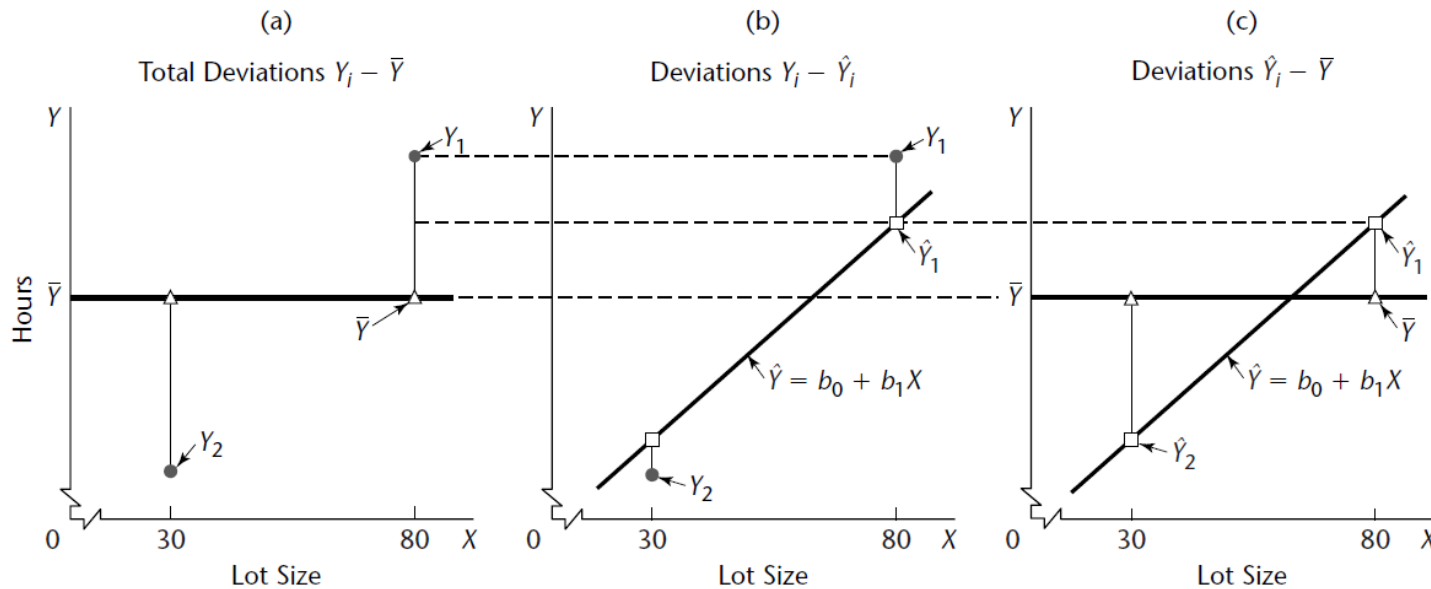
최소제곱 추정치

$$\mathbf{b}_{2 \times 1} = (\mathbf{X}'\mathbf{X})_{2 \times 2}^{-1} \mathbf{X}'\mathbf{Y}_{2 \times 1}$$

Unit 01 | 선형 회귀분석

제곱합 분해 (Partition of Sum of Squares)

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$ —Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



- ✓ 데이터 : Data on Lot Size & Work Hours, Toluca Company.
- ✓ 본 예시에서, (a)(=Total Deviations)는 (b)와 (c)의 합으로 분해될 수 있다.

이미지 출처 : Michael H. Kutner, Christopher J. Nachtsheim, John Neter,
<Applied Linear Regression Models>

Unit 01 | 선형 회귀분석

제곱합 분해 (Partition of Sum of Squares)

	자유도 (df)	제곱합 (SS)	제곱평균 (MS)	F
회귀 (Regression) SSR	p	SSR	MSR = SSR / p	F = MSR / MSE
잔차 (Residual) SSE	n-(p+1)	SSE	MSE = SSE / n-p-1	
총 (Total) SST	n-1	SST = SSR + SSE		

회귀식이 설명하지 못하는 부분.
작을수록 좋음!

Ex. 단순선형회귀분석(p=1)에서
잔차에 걸리는 제약조건은 2개!
(따라서 자유도는 n-2)

Contents

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가 지표

Unit 02 | 회귀진단

회귀진단 (Regression Diagnostics)

- 데이터가 회귀분석에 사용된 모형의 **가정**을 제대로 만족하고 있는지 확인하는 과정

[회귀모형 기본 가정]

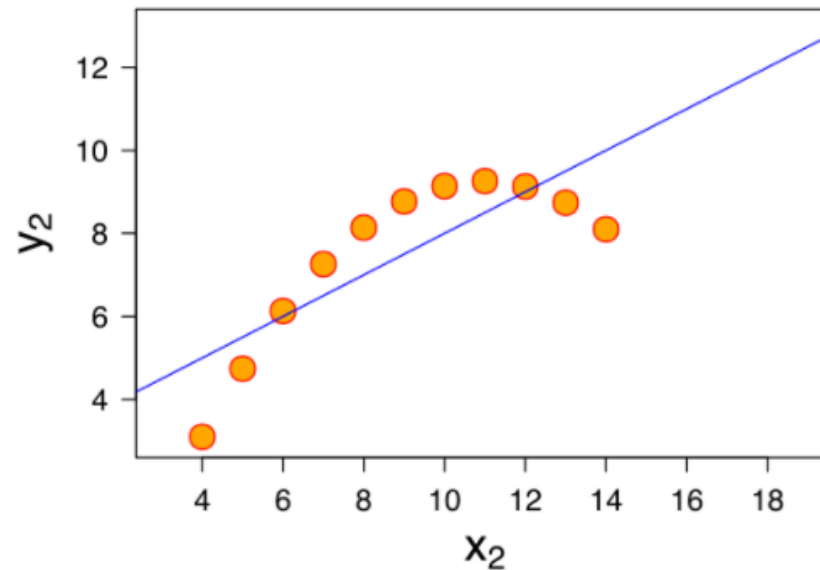
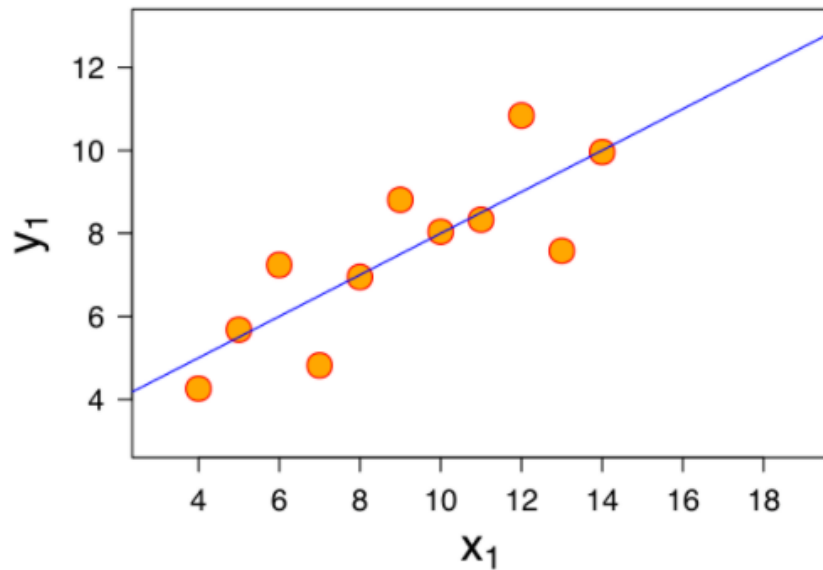
1. **선형성**(Linearity) : 설명변수(X)와 반응변수(Y) 간 선형 관계
2. **정규성**(Normality) : 오차(Error)의 정규성
3. **등분산성**(Homoscedasticity) : 오차의 등분산성
4. **독립성**(Independence) : 오차의 독립성



Unit 02 | 회귀진단

그래프적 방법

1. 선형성(설명변수와 반응변수 간 선형 관계) 판단

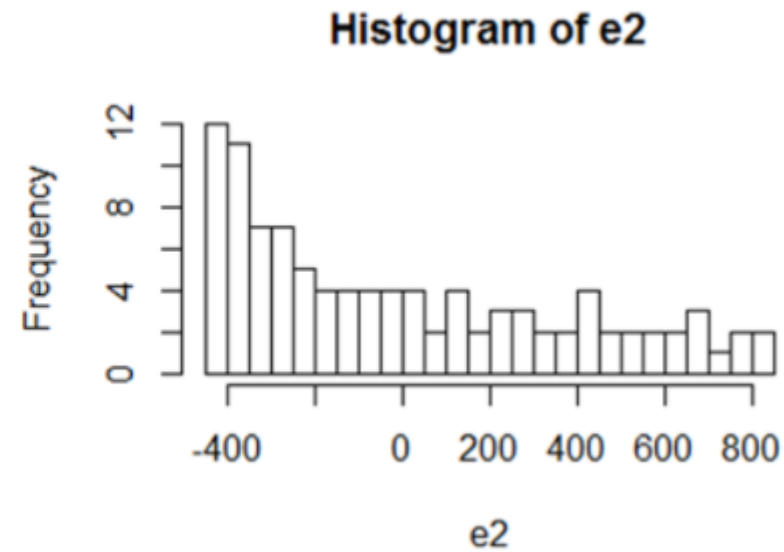
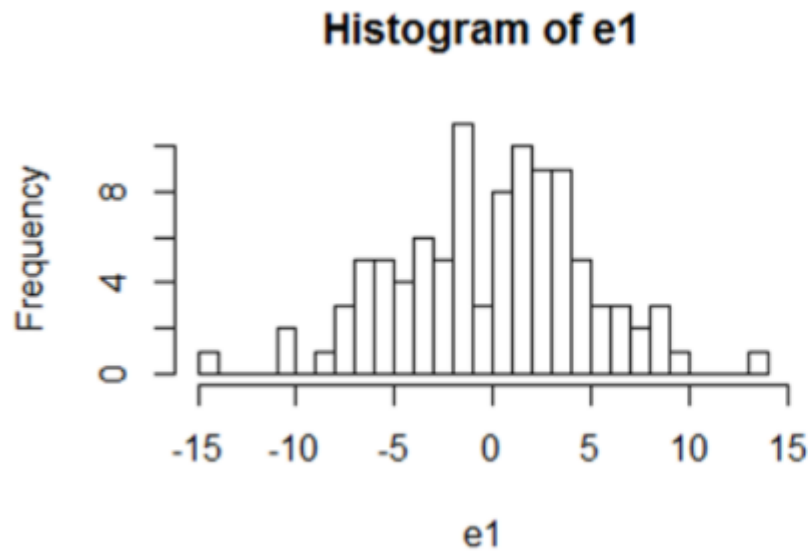


- 설명변수와 반응변수 간 산점도(Scatter plot)를 그려 선형성 판단 가능
- x_1 과 y_1 간에는 선형 관계가 존재하지만, x_2 와 y_2 간에는 선형 관계가 존재한다고 보기 어렵다.

Unit 02 | 회귀진단

그래프적 방법

2. 정규성(오차가 정규분포를 따르는지) 판단

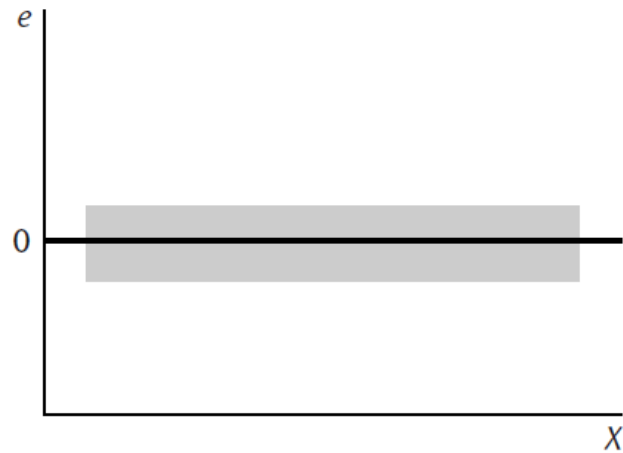


- 잔차의 히스토그램을 그려 오차의 정규성 판단 가능 cf. [R] Shapiro-Wilk Normality Test
- e1, e2의 분포로 보아 좌측은 정규성 가정을 만족하고, 우측은 정규성 가정을 위배한다고 추정

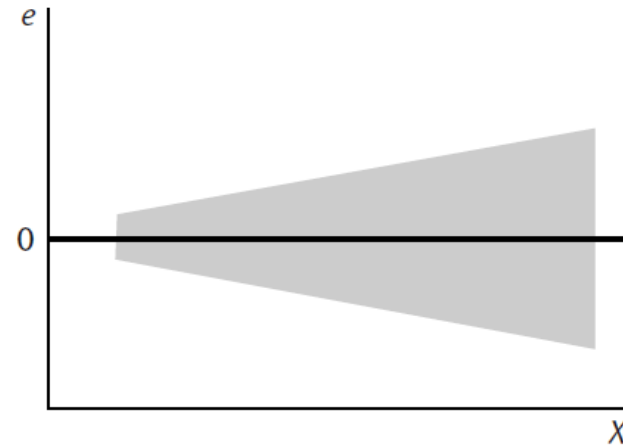
Unit 02 | 회귀진단

그래프적 방법

3. 등분산성(오차의 분산이 일정한지) 판단



(a)



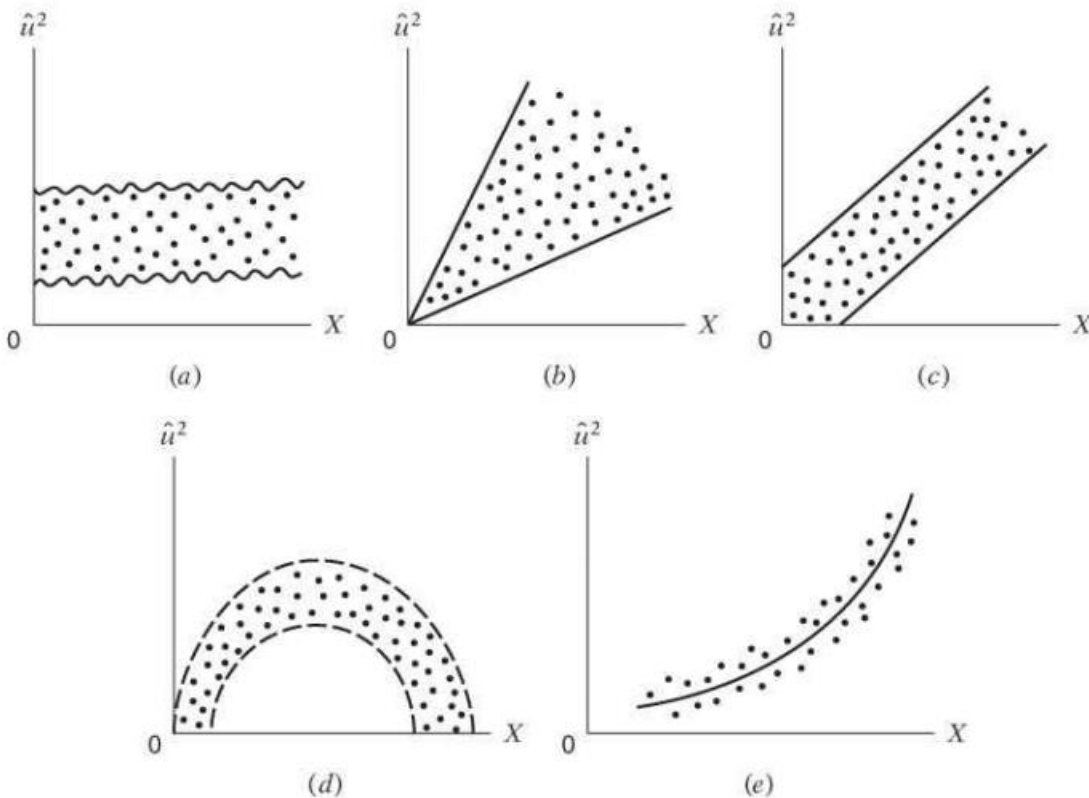
(c)

- 설명변수에 대한 잔차 그림(Plot of e against X)으로 오차의 등분산성 판단 가능
- (a)의 band width는 일정(등분산)하지만, (c)의 band width는 점점 커짐(이분산)

Unit 02 | 회귀진단

그래프적 방법

4. 독립성(오차가 서로 독립인지) 판단



- 설명변수와의 상관성, 자기 상관성 확인해서 독립성 판단 가능
- 직관적으로는, 잔차에 어떠한 패턴((b)~(e))이 있다면 독립적이지 않은 것!
- Durbin-Watson 검정, ACF

Unit 02 | 회귀진단

OLS : Ordinary Least Square

- Python **statsmodel** 패키지의 OLS 클래스 명령으로 선형 회귀분석 실시
- 모형 선택 기준**
 - R-squared & Adj. R-squared
 - F-statistic
 - t-statistic
 - Durbin-Watson (오차의 자기상관)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Mon, 18 Nov 2019	Prob (F-statistic):	6.72e-135
Time:	21:54:23	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

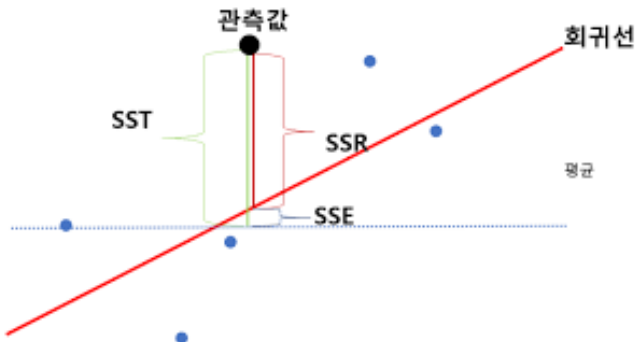
Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	1.51e+04

Unit 02 | 회귀진단

1. 결정계수(R-squared) & 조정된 결정계수(Adj. R-squared)

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

전체 제곱합(SSTO) 중 회귀식으로 설명(SSR) 가능한 부분
→ 따라서, 결정계수는 **클수록 좋다!**



SST(SSTO) : 총제곱합
SSR : 회귀제곱합
SSE : 잔차제곱합
SST = SSR + SSE

$$R_a^2 = 1 - \frac{\frac{SSE}{n-1}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SSTO}$$

결정계수의 **문제점** :
설명변수가 추가되면 SSR이 커지므로,
결정계수 값이 무조건 커짐!



조정된 결정계수 : **설명변수 개수에 대한 패널티** 부과

Unit 02 | 회귀진단

잠깐! 가설검정 Intro (1)

가설검정 : 모집단의 특징에 대한 통계적 가설을 표본을 통해 검토하는 방법

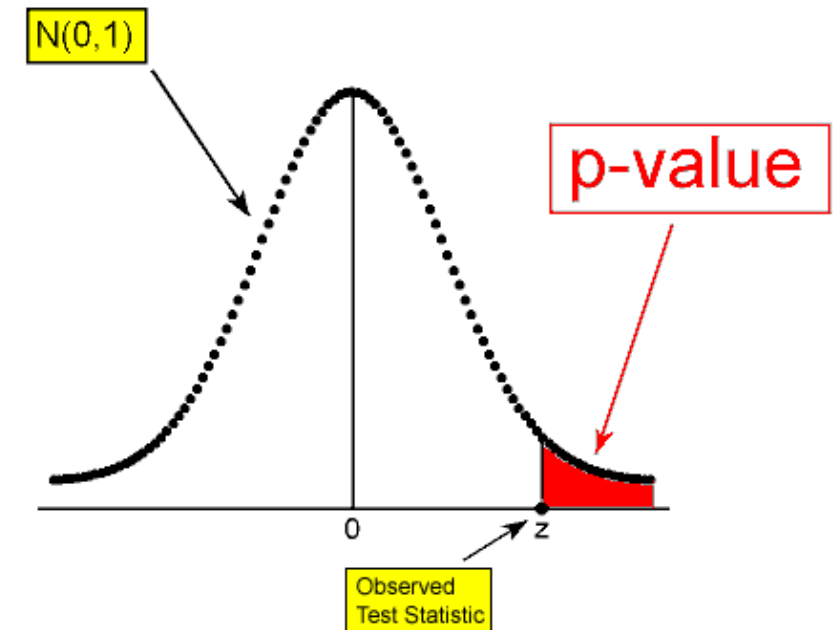
- 귀무가설(H_0 , Null hypothesis)
 - ✓ 모집단의 특징에 대해 옳다고 제안하는 잠정적 주장
 - ✓ “~와 차이가 없다.”, “~의 효과가 없다.” 등의 형태
 - ✓ 직접 검정 대상, 기각(reject)하는 것이 목표
- 대립가설(H_1 , Alternative hypothesis)
 - ✓ 귀무가설이 거짓이라면, 대안적으로 참이 되는 주장
 - ✓ “~와 차이가 있다.”, “~의 효과가 있다.” 등의 형태
 - ✓ 귀무가설이 기각되었을 때, 대안적으로 채택됨

Unit 02 | 회귀진단

잠깐! 가설검정 Intro (2)

가설검정 : 모집단의 특징에 대한 통계적 가설을 표본을 통해 검토하는 방법

- 검정통계량
 - ✓ 귀무가설이 옳다는 가정 하에 구해지는 통계량
 - ✓ 표본이 추출되어 계산된 통계량은 통계치
- p-value
 - ✓ 귀무가설이 옳다는 가정 하에 검정통계량이 관측될 확률



Unit 02 | 회귀진단

잠깐! 가설검정 Intro (3)

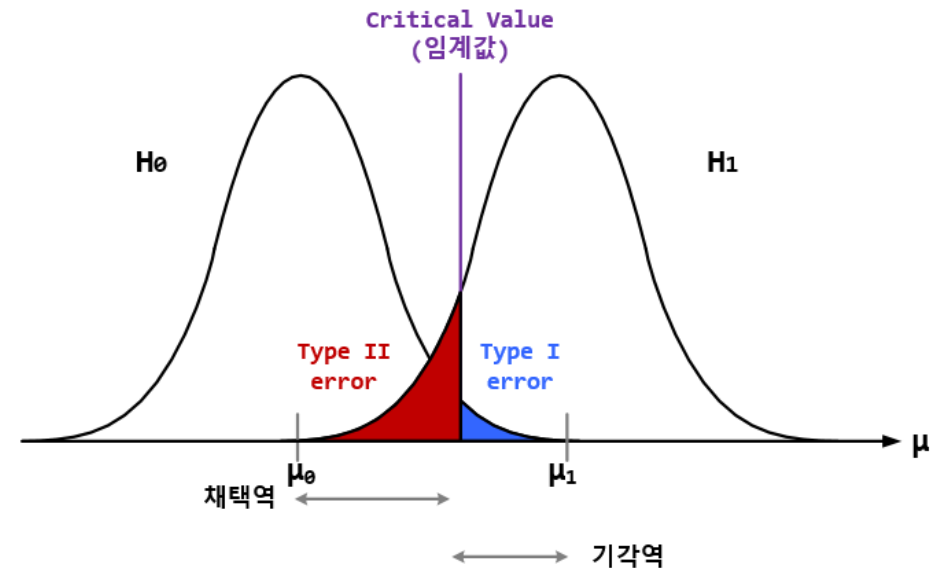
가설검정 : 모집단의 특징에 대한 통계적 가설을 표본을 통해 검토하는 방법

- 제1종오류

- ✓ 귀무가설이 참인 상황에서 귀무가설을 기각하는 오류
- ✓ 일반적으로, 제1종오류가 제2종오류보다 더 치명적
- ✓ 따라서 제1종오류의 **상한선** 설정 = **유의수준**

- 제2종오류

- ✓ 귀무가설이 거짓인 상황에서 귀무가설을 채택하는 오류



Unit 02 | 회귀진단

2. F-Statistic

OLS Regression Results

```
=====
Dep. Variable:          MEDV    R-squared:                0.741
Model:                  OLS     Adj. R-squared:             0.734
Method:                 Least Squares    F-statistic:             108.1
Date:                  Mon, 18 Nov 2019    Prob (F-statistic):       6.72e-135
Time:                  21:54:23    Log-Likelihood:          -1498.8
No. Observations:      506        AIC:                     3026.
Df Residuals:          492        BIC:                     3085.
Df Model:              13
Covariance Type:       nonrobust
=====
```

- 귀무가설(H_0) : $\beta_1 = \beta_2 = \dots = \beta_k = 0$ VS 대립가설(H_1) : $\beta_j \neq 0$, for some j
- 모형 자체의 유의미함을 판단하는 기준
- 모든 설명변수의 계수가 0인지, 하나라도 0이 아닌지를 판별

Unit 02 | 회귀진단

3. t-Statistic

	coef	std err	t	P> t	[0.025	0.975]
const	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

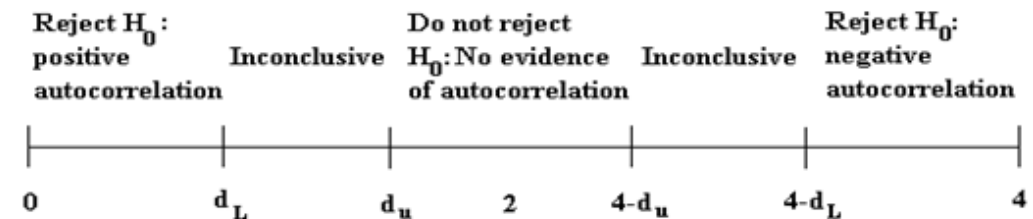
- 귀무가설(H_0) : $\beta_j = 0$ VS 대립가설(H_1) : $\beta_j \neq 0$
- 설명변수의 유의미함을 판단하는 기준
- 해당 설명변수의 계수가 0인지, 아닌지를 판별

Unit 02 | 회귀진단

4. Durbin-Watson (오차의 자기상관)

```
=====
Omnibus:                178.041    Durbin-Watson:                1.078
Prob(Omnibus):           0.000    Jarque-Bera (JB):           783.126
Skew:                    1.521    Prob(JB):                   8.84e-171
Kurtosis:                8.281    Cond. No.                   1.51e+04
=====
```

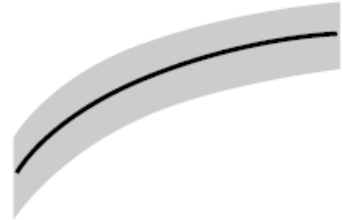

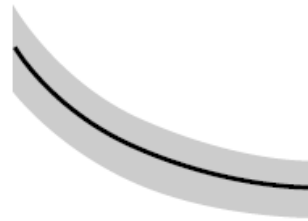
- Durbin-Watson : 오차의 독립성을 검정하기 위한 방법
- 0~4 범위의 값을 가지는데,
 - 0에 가깝다면 : 양의 상관관계
 - 4에 가깝다면 : 음의 상관관계
 - 2에 가깝다면 : 오차항의 자기상관 없음 (**독립성 만족**)



Unit 02 | 회귀진단

변수 변환 (Transformation)

- 변수 변환을 통해 회귀분석의 기본 가정을 충족시킬 수 있다.
- Transformation on X
 - X와 Y 간 비선형적 관계를 선형으로 변환 가능
 - 오차항의 spread는 변하지 않음
- Transformation on Y
 - X와 Y 간 비선형적 관계를 선형으로 변환 가능
 - 오차항의 spread가 변화함

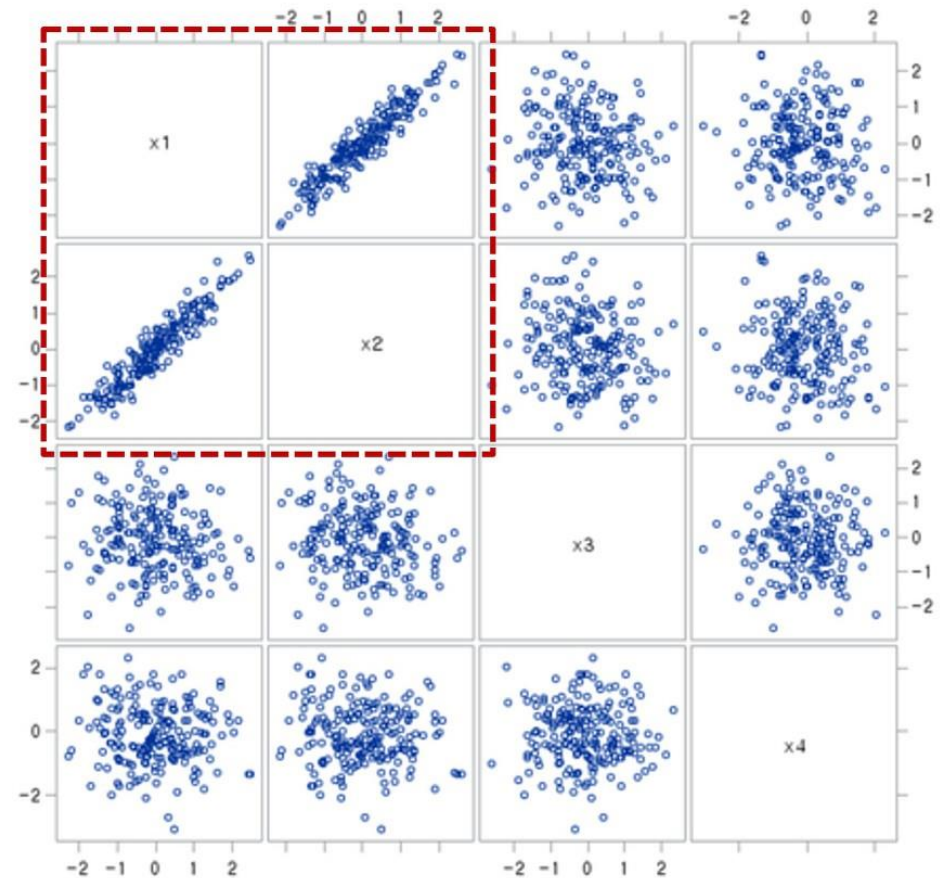
	Prototype Regression Pattern	Transformations of X
(a)		$X' = \log_{10} X$ $X' = \sqrt{X}$
(b)		$X' = X^2$ $X' = \exp(X)$
(c)		$X' = 1/X$ $X' = \exp(-X)$

Unit 02 | 회귀진단

다중공선성 (Multicollinearity)

cf. 회귀분석의 목적이 '개별 회귀계수의 추정' 이 아닌 '반응변수의 예측'이면, 다중공선성의 존재가 큰 문제가 되지는 않는다.

- 회귀분석에서 **설명변수들 간 강한 상관관계**('strongly correlated')가 나타나는 문제
- Detection
 - ✓ 설명변수 산점도, heatmap, 상관계수 행렬
- 다중공선성이 존재하면, **회귀계수의 추정이 불안정**
 - ✓ 분산이 매우 커져서 오류에 민감해짐
 - ✓ cf. [선형대수] 모든 column이 선형 독립(linearly independent)이어야, 하나 이상의 해가 존재



Unit 02 | 회귀진단

다중공선성의 진단 – VIF (Variance Inflation Factor)

- VIF(Variance inflation factor)

$$VIF_i = \frac{1}{1 - R_i^2}$$

VIF가 10 이상인 경우 다중공선성이 있는 변수라고 판단

$$x_1 = \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon$$

X1을 종속변수, 나머지 변수를 독립변수로 하여 회귀 모델(f_1) 적합

$$R_1^2$$

f_1 의 R^2 를 이용하여 VIF_1 계산

$$VIF_1 = \frac{1}{1 - R_1^2}$$

VIF_1 의 의미 : 다른 변수의 선형결합으로 X1을 설명할 수 있는 정도

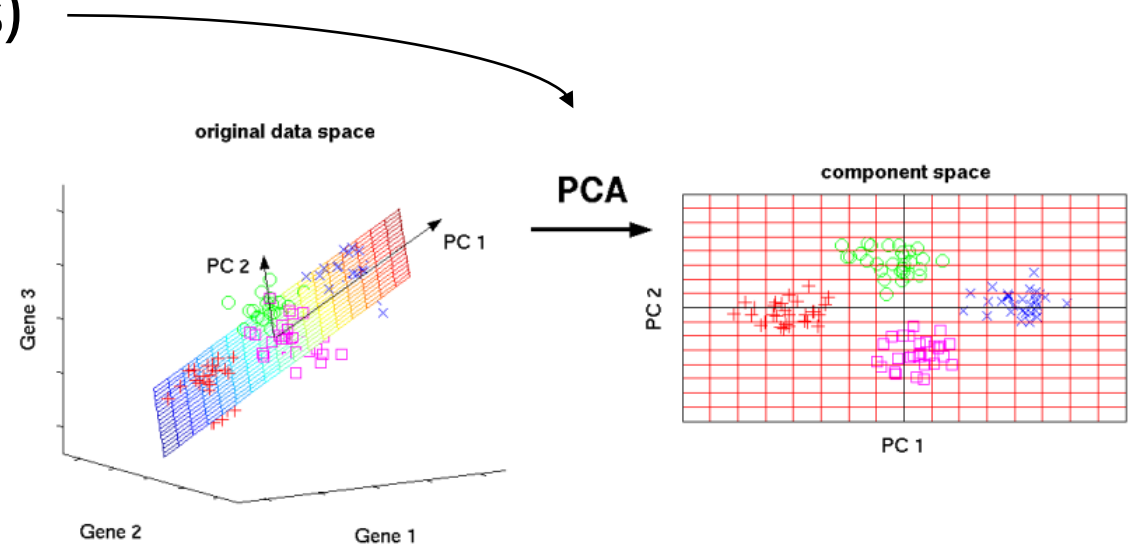
$R^2 > 0.9$ 이상인 경우, $VIF > 10$

- 일반적으로 VIF 값이 10보다 크다면, 다중공선성이 존재한다고 판단!
- 단, VIF 값이 크더라도 해당 설명변수가 통계적으로 유의하다면 제거하지 않는 편이 바람직하다.

Unit 02 | 회귀진단

다중공선성의 제거

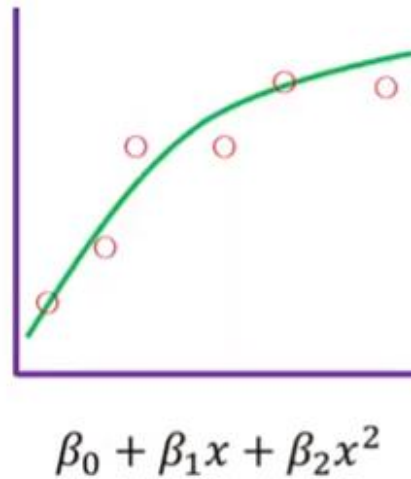
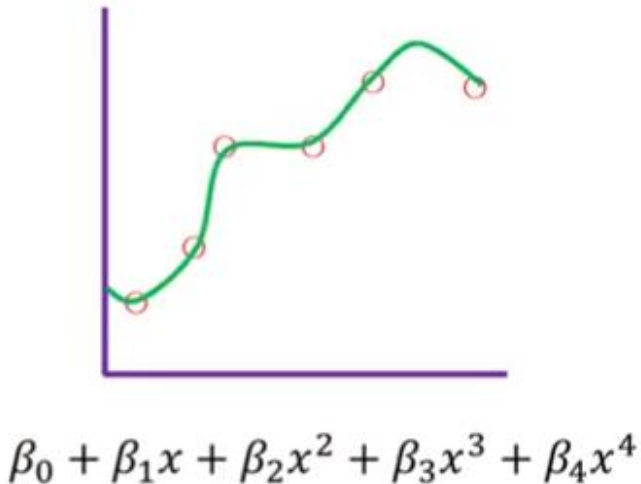
- 설명변수 제거
 1. 다중공선성을 유발하는 설명변수 2개를 찾아, 각 변수를 제거했을 때 R-squared의 변동 확인
 2. 제거했을 때 R-squared 값이 유지되는 설명변수 제거
- PCA(주성분 분석, Principal Component Analysis)
- Ridge / Lasso Regression (일종의 패널티 부과)



Unit 02 | 회귀진단

다중공선성의 제거 - Ridge Regression(L2 Regression)

- 정규화(Regularization)를 이용한 대표적인 shrinkage 방법



- 좌측 모델은 '과적합(Overfitting)'
 - ✓ 현재 데이터(train data)를 완벽하게 설명하고 있지만,
 - ✓ 미래 데이터(test data)에 대한 예측력은 떨어질 것
- 따라서, 우측 모델을 사용하는 편이 바람직
- 정규화 : 과적합 모델이 일반성을 갖추도록!

Unit 02 | 회귀진단

다중공선성의 제거 – Ridge Regression(L2 Regression)

- 정규화(Regularization)를 이용한 대표적인 shrinkage 방법

$$\beta_1, \beta_2, \dots, \beta_p$$

$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{(1) \text{ Training accuracy}} + \lambda \underbrace{\sum_{j=1}^p \beta_j^2}_{(2) \text{ Generalization accuracy}}$$

- Training Accuracy에 Generalization Accuracy를 추가
 - 회귀계수(β)에 제약을 줄 수 있게 됨
 - 이렇게 계수 추정치를 줄여주는 정규화 방법이 'shrinkage'

Unit 02 | 회귀진단

다중공선성의 제거 – Ridge Regression(L2 Regression)

- **정규화**(Regularization)를 이용한 대표적인 **shrinkage** 방법
- Ridge Regression은 정규화 컨셉을 처음 도입한 모델
 - 기존 모델을 정규화하여, 조금 더 좋은 performance를 낼 수 있는 기초 기법

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$$



$$RSS + \lambda \sum_{j=1}^p \beta_j^2$$

Unit 02 | 회귀진단

다중공선성의 제거 – Lasso Regression(L1 Regression)

- Ridge Regression과 유사하나, **패널티 항으로 회귀계수의 절댓값 합을 이용**
- Feature Selection과 관련 → 중요한 설명변수 몇 개만 추리고, 나머지는 0으로!

$$RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$$



$$RSS + \lambda \sum_{j=1}^p |\beta_j|$$

Unit 02 | 회귀진단

다중공선성의 제거 – Elastic Net Regression

- Ridge Regression과 Lasso Regression의 절충
- 큰 데이터셋에서 잘 작동하며, L2, L1 norm에 대한 가중치를 조절해가면서 사용
 - 교차검증(Cross Validation)을 통해 하이퍼파라미터 튜닝
- [R] glmnet 패키지, glmnet() 함수

$$\operatorname{argmin}_{\beta_j} \left(\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^n \left(\alpha |\beta_j| + (1 - \alpha) \beta_j^2 \right) \right) \quad (0 \leq \alpha \leq 1)$$

Unit 02 | 회귀진단

선형 회귀분석 마무리

1. 회귀모형 설정 : 반응변수 및 주요 설명변수 파악
2. 선형성 검토 : 산점도를 통해 상관관계 파악
3. 설명변수 검토 : 각 설명변수 분포 확인 및 다중공선성 점검
4. 모델 적합 : 모델 회귀계수 추정 및 모형 적절성 검토
5. 변수 선택 : 주요 설명변수 선택
6. 적합한 모형 검토 : 오차에 대한 기본 가정 확인
7. 최종 모형 선택

Contents

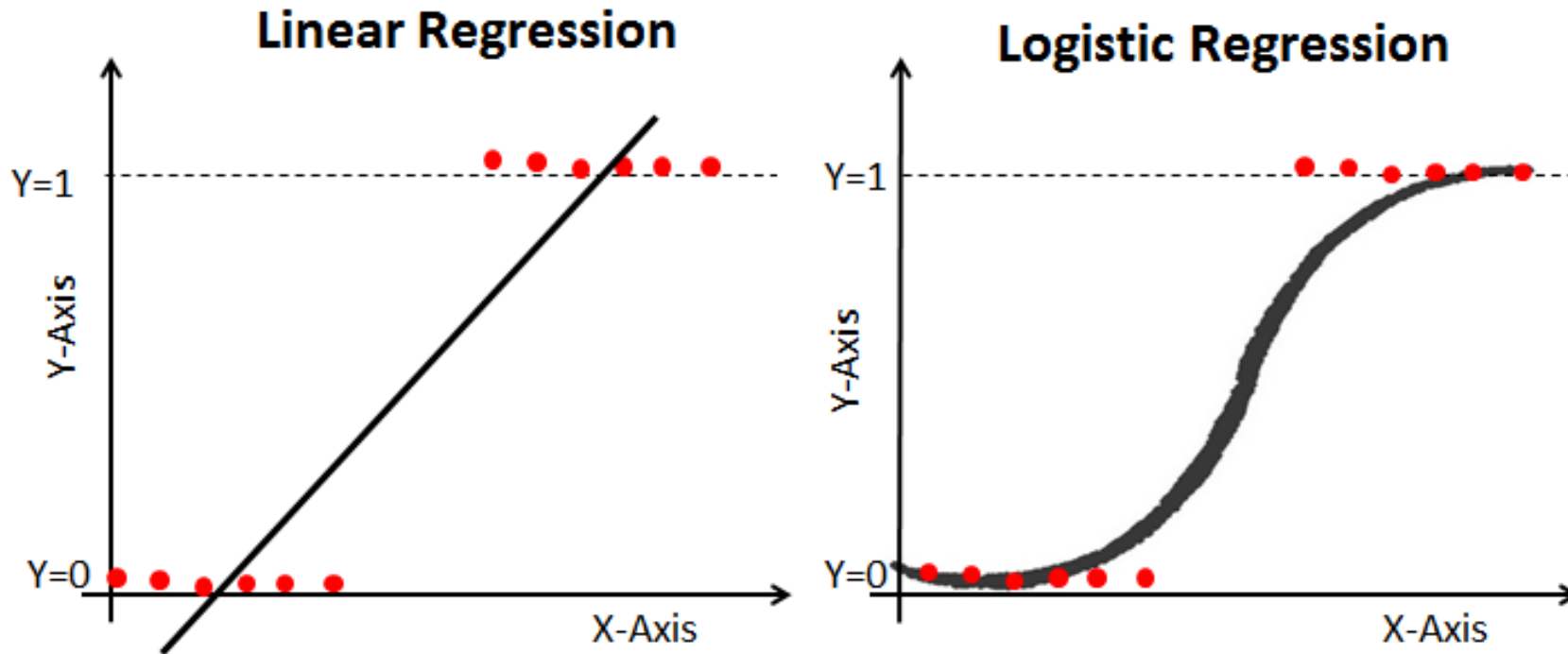
Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가 지표

Unit 03 | 로지스틱 회귀분석

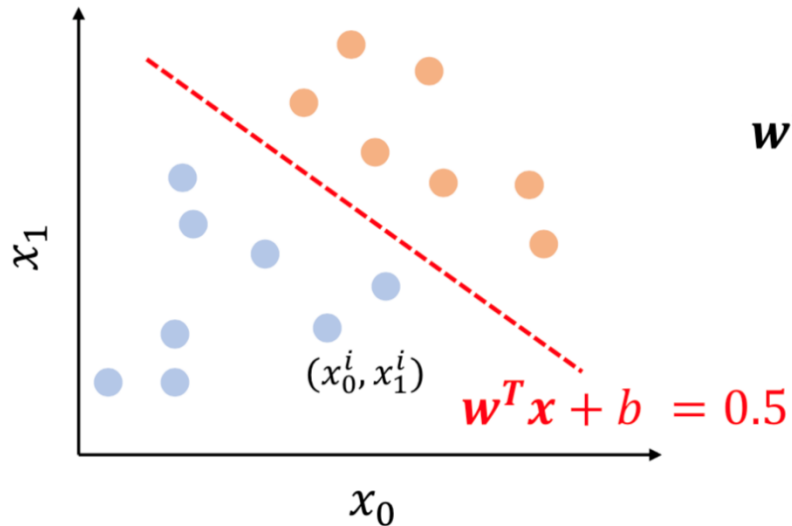


- 반응변수 Y가 위와 같이 범주형(categorical) 변수라면, 선형 회귀분석은 적절하지 않다.
 - 이 경우, **로지스틱 회귀분석**으로 주어진 데이터를 **분류**할 수 있다.

Unit 03 | 로지스틱 회귀분석

로지스틱 회귀분석 (Logistic Regression)

- 범주형 데이터를 대상으로 하는 회귀분석, 일종의 **분류**(Classification) 기법
 - Ex. 제품 불량 여부(양품/불량), 고객 이탈 여부(이탈/잔류), 정상 거래 여부(정상/사기) 등



$$w^* = \operatorname{argmin}_w \{-\sum_i y \log \tilde{y} + (1 - y) \log(1 - \tilde{y})\}$$

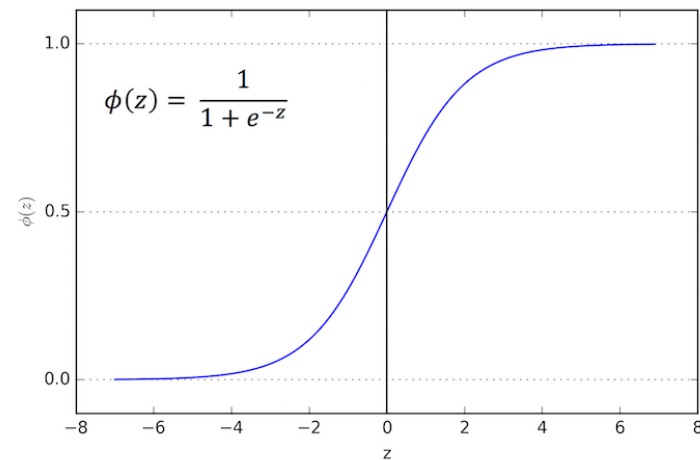
“이진 교차 엔트로피를 최소화한다”

로지스틱 회귀는 선형 회귀와 비슷하나, 범주형 데이터를 분류하는 방향으로 선을 긋는다.

Unit 03 | 로지스틱 회귀분석

로지스틱 함수 (Logistic Function)

- 시그모이드(Sigmoid) 함수 : 실수 전체를 정의역으로 하고, 유한한 범위 내에서 단조 증가
 - ✓ 딥러닝 활성화 함수로 활용
- **로지스틱 함수** : 음의 무한대($-\infty$)부터 양의 무한대(∞)까지의 실수값을 0부터 1 사이의 실수값으로 대응시키는 시그모이드 함수
 - ✓ Output 범위 : (0, 1) / Input에 대해 단조 증가

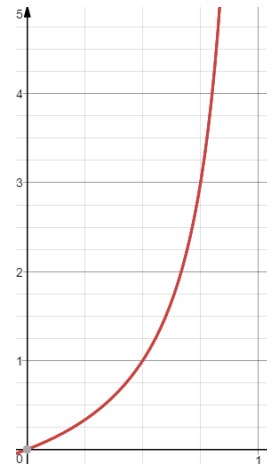


Unit 03 | 로지스틱 회귀분석

로지스틱 함수 식 - Odds ratio에 대한 Logit Transformation

- 승산비(Odds ratio) : 실패 확률 대비 성공 확률 비율

$$\text{odds ratio} = \frac{\mu}{1 - \mu}$$



- 로짓 변환(Logit Transformation) : Odds ratio에 log를 취하는 변환으로, 입력 값의 범위가 [0, 1] 일 때 출력 값의 범위를 $(-\infty, \infty)$ 로 조정

$$z = \text{logit}(\text{odds ratio}) = \log\left(\frac{\mu}{1 - \mu}\right)$$

Unit 03 | 로지스틱 회귀분석

선형 판별함수

$$\text{logistic}(z) = \mu(z) = \frac{1}{1 + \exp(-z)}$$

- 로지스틱 함수를 사용하는 경우, z 값과 μ 값 간에는 다음과 같은 관계가 성립한다.
 - $z = 0$ 일 때 $\mu = 0.5$
 - $z > 0$ 일 때 $\mu > 0.5 \rightarrow \hat{y} = 1$
 - $z < 0$ 일 때 $\mu < 0.5 \rightarrow \hat{y} = 0$

즉, z 가 분류 모형의 판별함수(decision function)의 역할을 한다고 볼 수 있다.

Unit 03 | 로지스틱 회귀분석

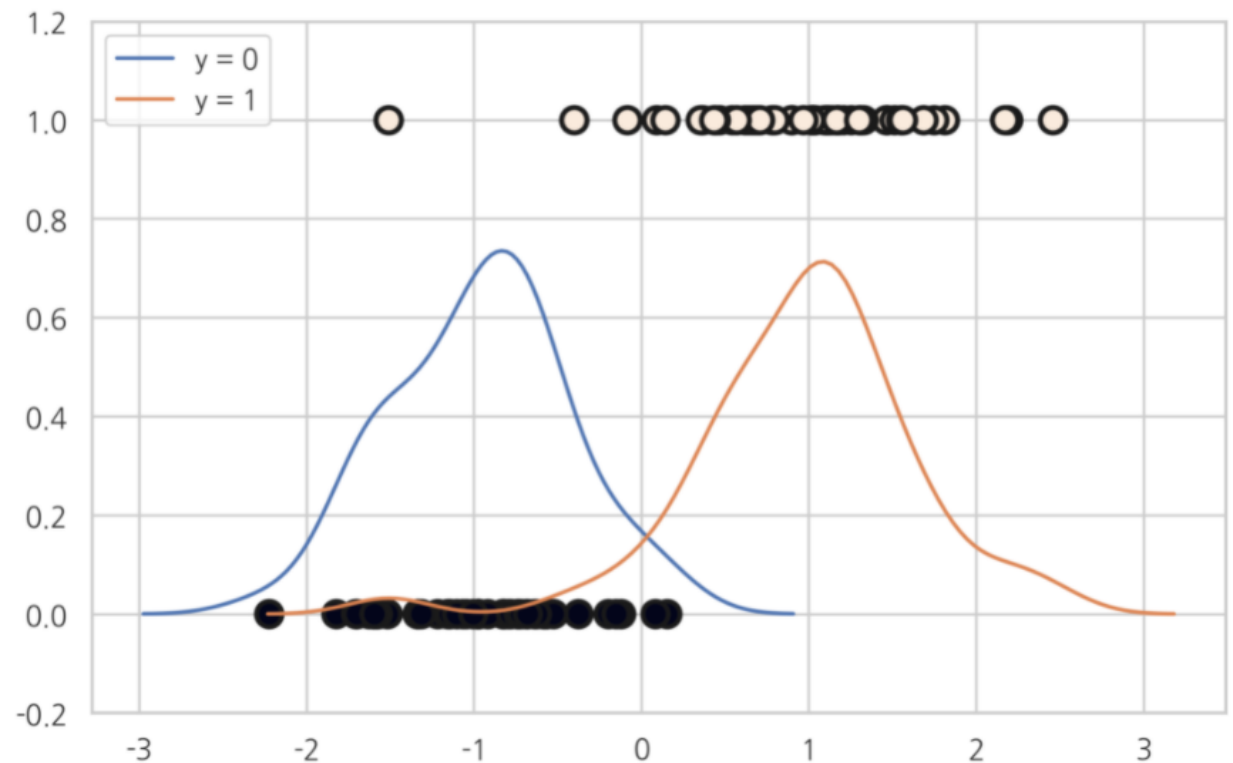
로지스틱 회귀분석 예제

- 1차원 설명변수를 가지는 분류문제
- 파이썬 StatsModels 패키지 활용

```
from sklearn.datasets import make_classification

X0, y = make_classification(n_features=1, n_redundant=0, n_informative=1,
                           n_clusters_per_class=1, random_state=4)

plt.scatter(X0, y, c=y, s=100, edgecolor="k", linewidth=2)
sns.distplot(X0[y == 0, :], label="y = 0", hist=False)
sns.distplot(X0[y == 1, :], label="y = 1", hist=False)
plt.ylim(-0.2, 1.2)
plt.show()
```



Unit 03 | 로지스틱 회귀분석

로지스틱 회귀분석 예제

- StatsModels 패키지에서는 베르누이 분포를 따르는 로지스틱 회귀모형 Logit을 제공
- 사용 방법은 OLS 클래스와 동일
- summary 메소드로 리포트 출력

```
X = sm.add_constant(X0)
logit_mod = sm.Logit(y, X)
logit_res = logit_mod.fit(dis=0)
print(logit_res.summary())
```

Logit Regression Results

Dep. Variable:	y	No. Observations:	100
Model:	Logit	Df Residuals:	98
Method:	MLE	Df Model:	1
Date:	Sat, 06 Jun 2020	Pseudo R-squ.:	0.7679
Time:	10:01:05	Log-Likelihood:	-16.084
converged:	True	LL-Null:	-69.295
Covariance Type:	nonrobust	LLR p-value:	5.963e-25

	coef	std err	z	P> z	[0.025	0.975]
const	0.2515	0.477	0.527	0.598	-0.683	1.186
x1	4.2382	0.902	4.699	0.000	2.470	6.006

Unit 03 | 로지스틱 회귀분석

회귀계수의 해석

- 선형 회귀: 설명변수가 1만큼 증가할 때 **반응변수**의 변화량
- 로지스틱 회귀 : 설명변수가 1만큼 증가함에 따른 **$\log(\text{Odds})$** 의 변화량

Contents

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가 지표

Unit 04 | 최대우도추정 & 평가 지표

회귀계수의 추정

- 선형 회귀분석 → 최소제곱합(LSE) 이용
- 로지스틱 회귀분석 → 최대 우도 추정법(MLE) 이용

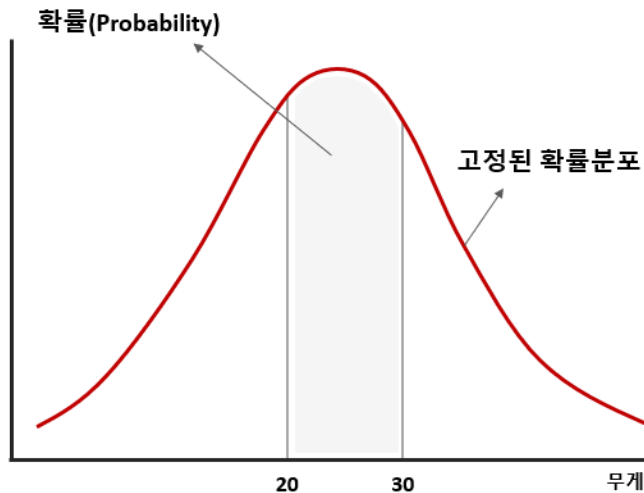
↓

최대 우도 추정법 (Maximum Likelihood Estimation)

- 회귀식이 비선형이므로, LSE 사용 불가
- Likelihood를 Maximize하는 parameter 추정

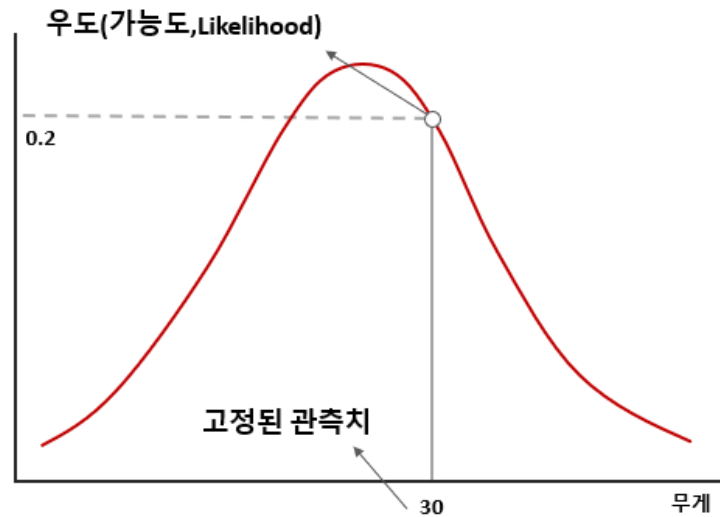
Unit 04 | 최대우도추정 & 평가 지표

Probability(확률) VS Likelihood(우도, 가능도)



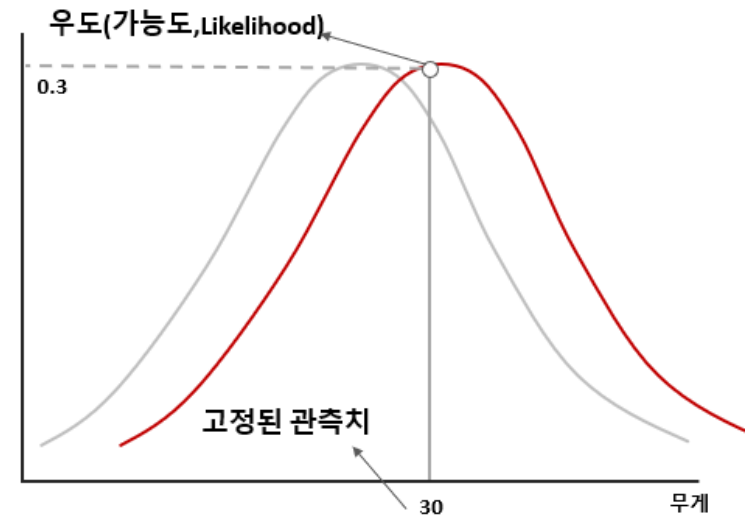
Probability(확률)

- 주어진(고정된) 확률 분포에서, 특정 관측값이 나타날 가능성



Likelihood(우도, 가능도)

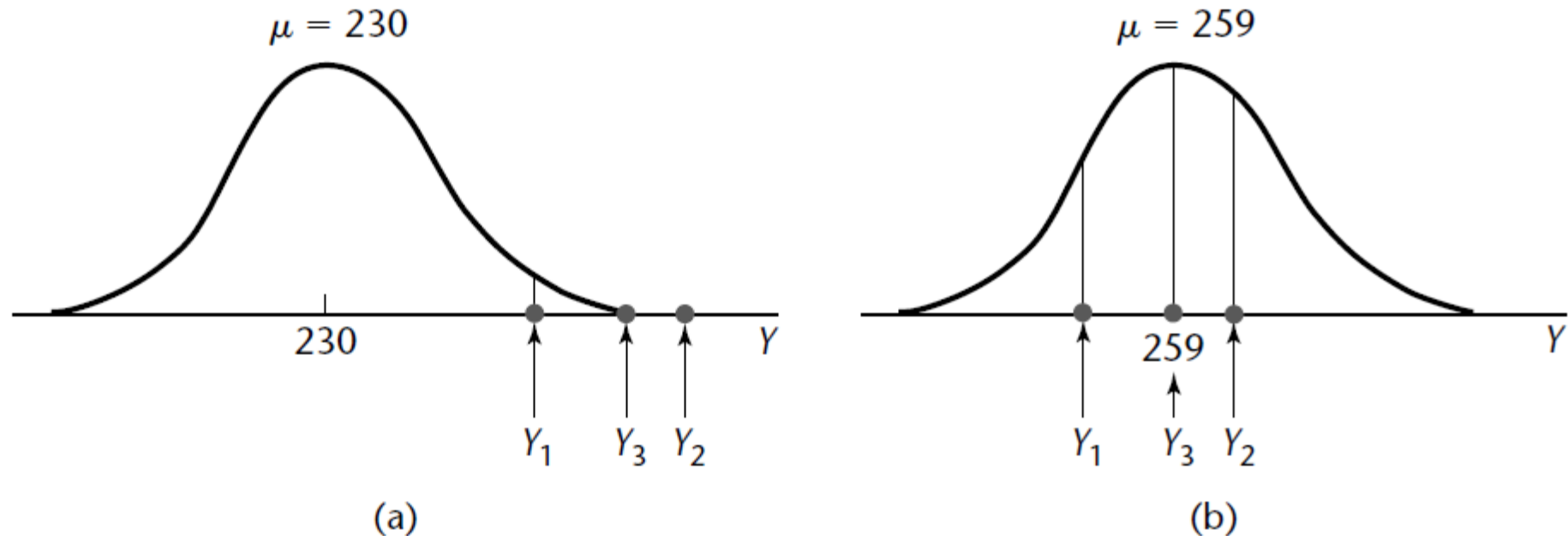
- 주어진(고정된) 관측값이, 특정 확률 분포에서 나타날 가능성
- 즉, 데이터가 특정 분포로부터 생성(generate)될 확률



Unit 04 | 최대우도추정 & 평가 지표

MLE 직관적으로 이해하기

FIGURE 1.13
Densities for
Sample
Observations
for Two
Possible Values
of μ : $Y_1 = 250$,
 $Y_2 = 265$,
 $Y_3 = 259$.



- 세 개의 관측값 Y_1, Y_2, Y_3 은 (a) 분포보다는 (b) 분포로부터 생성되었을 것이라는 예측
- True parameter는 259(분포 (b)의 평균)일 가능성이 높아 보임

Unit 04 | 최대우도추정 & 평가 지표

최대 우도 추정량(Maximum Likelihood Estimator) 찾기

$$L(\theta) = p(X|\theta) = \prod_{n=1}^N p(x_n|\theta)$$



수리적 편의(미분 용이)를 위해 양변에 log를 취한 후, -를 붙인
Negative log likelihood (목표 : minimize)

$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^N \ln p(x_n|\theta)$$



아래 식을 만족하는(편미분 값이 0이 되는) 모수 찾기

$$\frac{\partial}{\partial \theta} E(\theta) = -\frac{\partial}{\partial \theta} \sum_{n=1}^N \ln p(x_n|\theta) = -\sum_{n=1}^N \frac{\frac{\partial}{\partial \theta} p(x_n|\theta)}{p(x_n|\theta)} \stackrel{!}{=} 0$$

Unit 04 | 최대우도추정 & 평가 지표

예제 - MLE for Normal Error Regression Model

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right] \end{aligned}$$

Parameter	Maximum Likelihood Estimator
β_0	$\hat{\beta}_0 = b_0$ same as (1.10b)
β_1	$\hat{\beta}_1 = b_1$ same as (1.10a)
σ^2	$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n}$

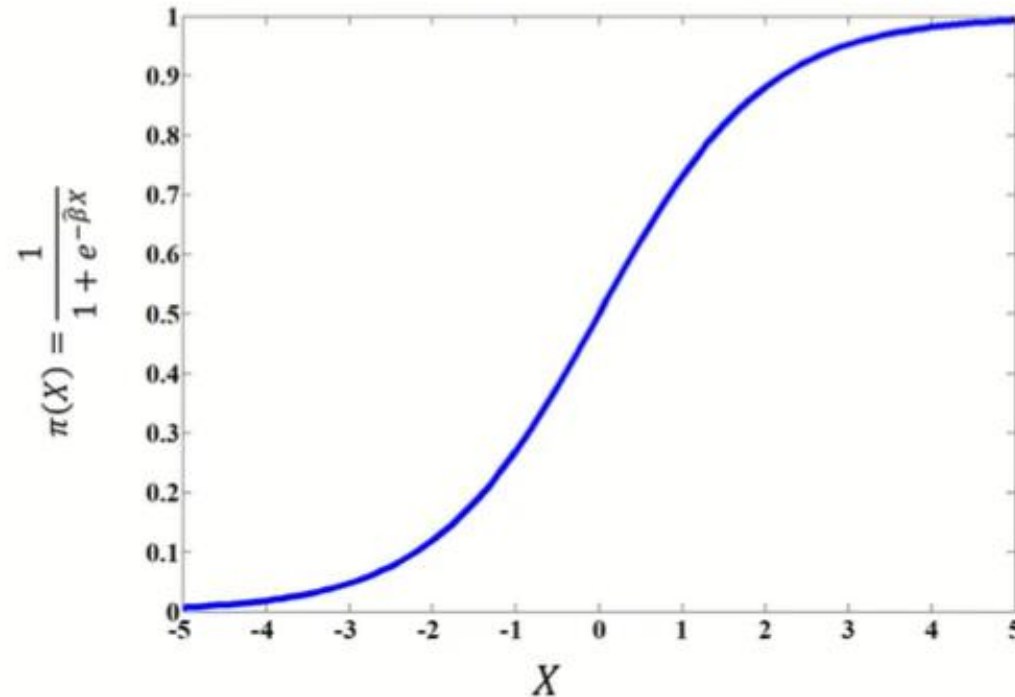
회귀계수의 경우 LSE의 결과와 동일하며,
분산은 LSE의 결과와 근소한 차이

$$s^2 = MSE = \frac{n}{n-2} \hat{\sigma}^2$$

Unit 04 | 최대우도추정 & 평가 지표

최종 로지스틱 회귀모델 - 최적의 parameter 적합

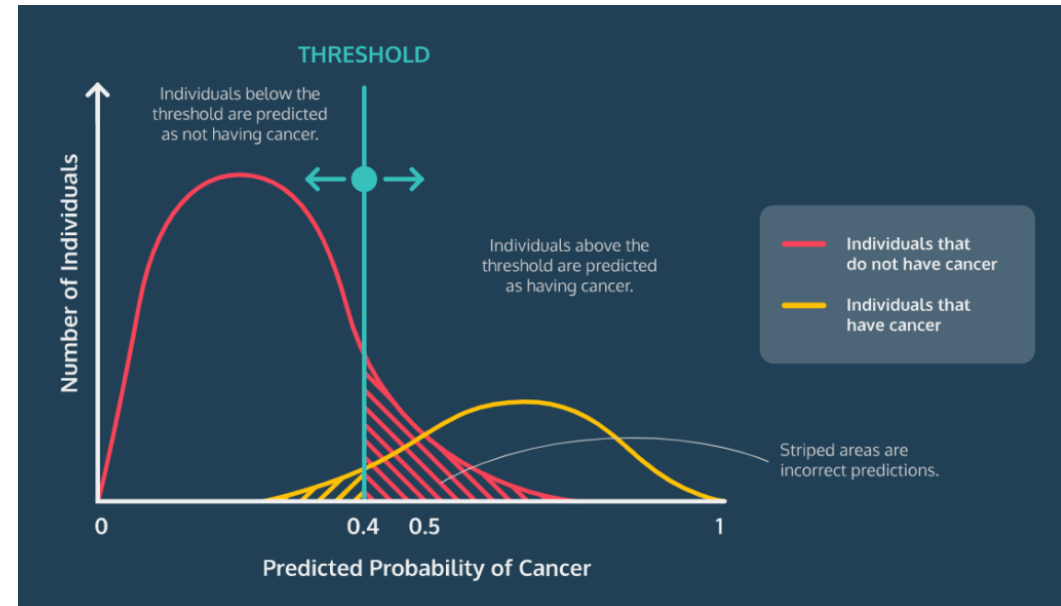
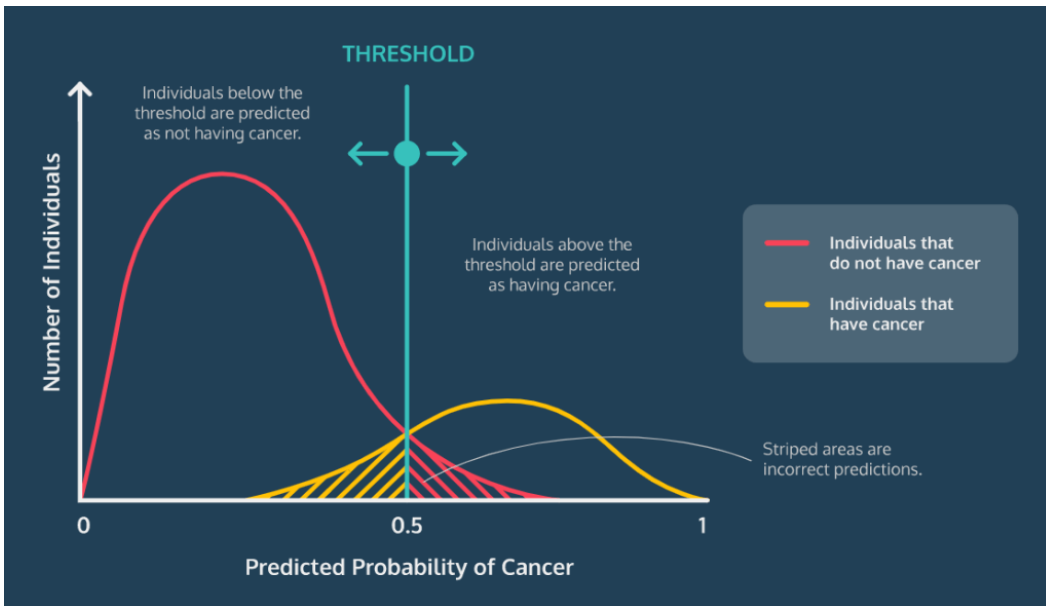
$$\pi(X) = f(X) = \frac{1}{1 + e^{-(\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \dots + \widehat{\beta}_p X_p)}} = \frac{1}{1 + e^{-\widehat{\beta}X}}$$



Unit 04 | 최대우도추정 & 평가 지표

Cutoff(Threshold)

- 분류(Classification)를 위한 기준
- 로지스틱 함수로 구한 확률이 cutoff 이상이면 1, cutoff 이하이면 0으로 분류
- Cutoff을 조정하여 성능 조절 가능



Unit 04 | 최대우도추정 & 평가 지표

성능 평가 지표 - 1. 정밀도 (Precision)

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

정밀도 (Precision)

- 모델이 True로 분류한 것 중 실제 True인 것의 비율
- $$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- PPV(Positive Predictive Value)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

Unit 04 | 최대우도추정 & 평가 지표

성능 평가 지표 - 2. 재현율 (Recall)

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

재현율 (Recall)

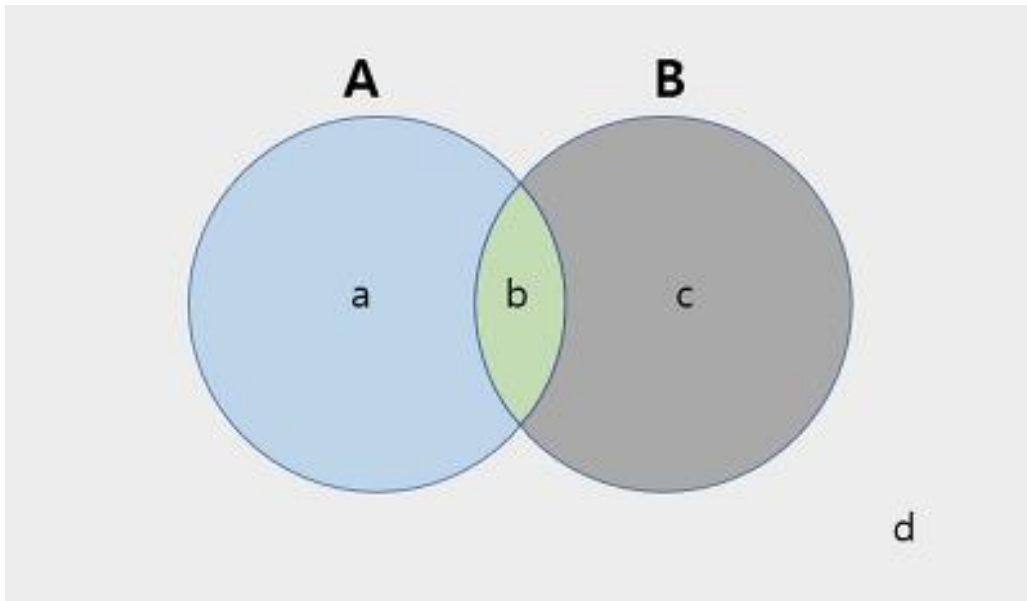
- 실제 True인 것 중 모델이 True로 분류한 것의 비율
- $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
- 통계학에서는 'Sensitivity'

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

Unit 04 | 최대우도추정 & 평가 지표

성능 평가 지표 - 정밀도(Precision)와 재현율(Recall)

Ex. 날씨 예측(맑다/흐리다) 모델

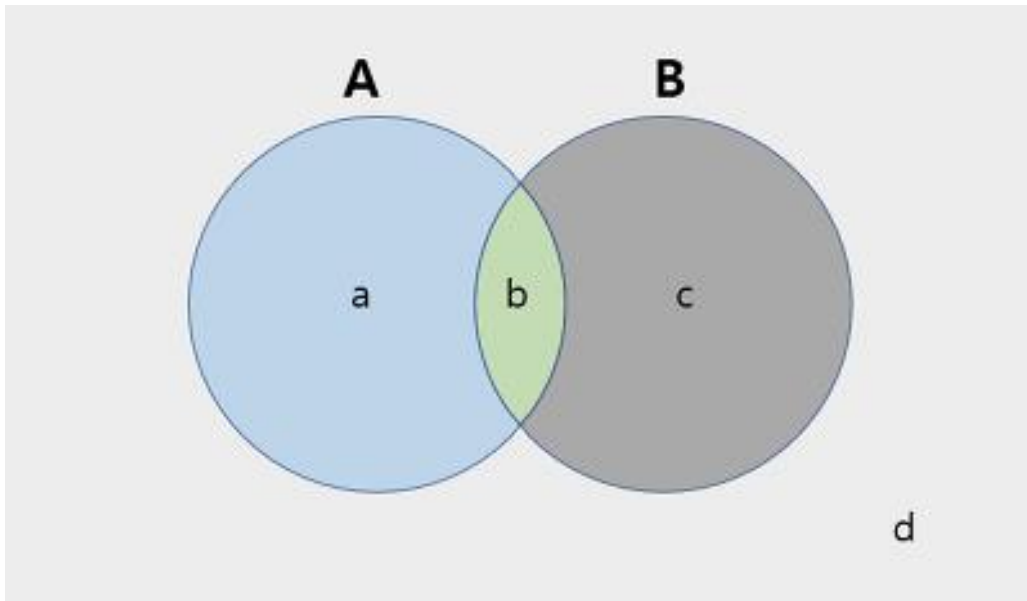


- A : 실제 날씨가 맑은 날
- B : 모델이 날씨가 맑다고 예측(분류)한 날
- $b = TP$ = 실제 날씨가 맑은 날을 모델이 날씨가 맑다고 예측(제대로 예측)한 날
- 이때,
 - ✓ $\text{Precision} = \frac{b}{b+c}$
 - ✓ $\text{Recall} = \frac{b}{a+b}$
 - ✓ a의 영역이 줄어들면 c의 영역이 커지게 됨
= 두 지표 간 Trade-off 관계

Unit 04 | 최대우도추정 & 평가 지표

Precision과 Recall 간 Trade-off 관계를 표로 보면,

Ex. 날씨 예측(맑다/흐리다) 모델



		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(40)
	False	FN(30)	TN(10)

Precision = 33.3%
Recall = 40%

		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(80)

Precision = 20%
Recall = 100%

Unit 04 | 최대우도추정 & 평가 지표

성능 평가 지표 - 3. Accuracy

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

정확도 (Accuracy)

- 예측 결과가 실제와 얼마나 동일한지 측정
- $$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$
- 가장 직관적으로 모델 성능 예측 가능
- 데이터 분포가 skewed(도메인 불균형) → 적합 X

Unit 04 | 최대우도추정 & 평가 지표

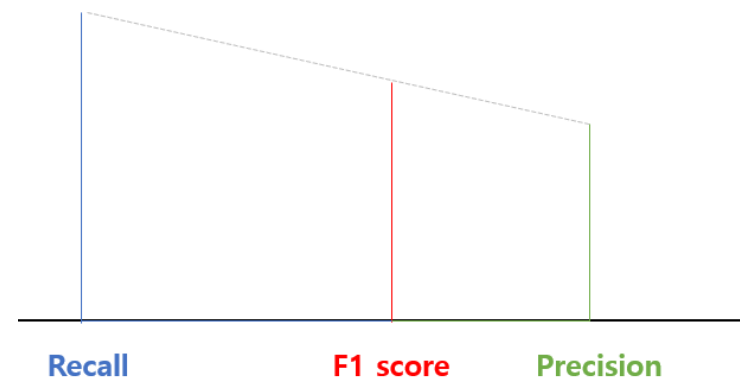
성능 평가 지표 - 4. F1 Score

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

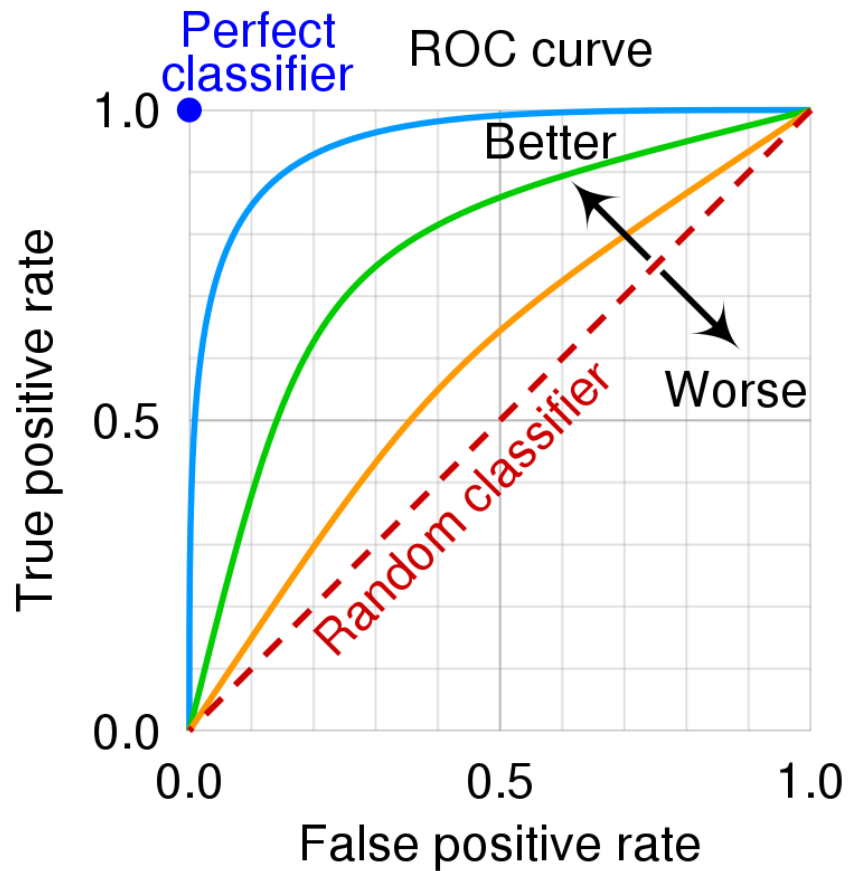
F1 Score



- Precision과 Recall의 조화 평균
- $$F1\ Score = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$
- 산술평균보다 큰 비중이 끼치는 bias ↓

Unit 04 | 최대우도추정 & 평가 지표

성능 평가 지표 - 5. ROC Curve



ROC Curve

- Confusion Matrix에서 FPR, Recall(Sensitivity) 값 계산
 - $FPR = \frac{FP}{FP+TN}$ (False Positive Rate)
- **AUC(=Area Under Curve)** : ROC Curve 아래 면적
 - ✓ 최댓값은 1
 - ✓ 값이 클수록(FPR에 비해 Recall이 클수록) 모델 성능 좋음

Unit 04 | 최대우도추정 & 평가 지표

로지스틱 회귀분석 마무리

1. 범주형 반응변수 Y 분류를 위한 기법
2. 로지스틱 함수의 출력값은 0과 1 사이
3. Logit 함수 : $\log(\text{Odds}) = \log(p/(1-p))$
4. Beta1 : $\log(\text{Odds})$ 의 변화량
5. 최대 우도 추정법(MLE)으로 최적의 parameter 찾기
6. Cutoff value 조정을 통해 분류 성능 조정 가능

과제

[과제 1]

- LSE 정규방정식, MSE 구현

[과제 2] 회귀분석 - Used Car Priced Prediction

- Ch 1, Ch 2를 토대로 자유롭게 회귀분석 & 회귀진단 진행
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

[과제 3] 로지스틱 회귀분석 - Credit Card Fraud Detection

- 파이썬 sklearn 패키지를 활용해 로지스틱 회귀분석 진행
- 성능지표 계산 및 해석
 - sklearn의 mean accuracy, f1 score 등
 - confusion matrix의 tp, fp, fn, tn 값
- 성능 개선 시도 (어떤 성능지표를 기준으로 했는지, 해당 지표 선택 이유 등)
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

Reference

[강의안]

- 투빅스 14기 강재영님 강의안
- 투빅스 15기 장아연님 강의안
- 연세대학교 응용통계학과 김현태 교수님 <회귀분석> 강의안

[교재]

- Michael H. Kutner, Christopher J. Nachtsheim, John Neter, <Applied Linear Regression Models>

[참고 자료]

- [선형, 로지스틱] 데이터 사이언스 스쿨 4장, 6장 (<https://datascienceschool.net/intro.html>)
- [로지스틱] <https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/>
- [Ridge/Lasso Regression] [Ridge regression\(능형 회귀\) 간단한 설명과 장점 \(tistory.com\)](https://tistory.com)
- [회귀진단] [Regression\(03\) - 회귀진단 | DataLatte's IT Blog \(heung-bae-lee.github.io\)](https://heung-bae-lee.github.io)

Q & A

들어주셔서 감사합니다.