15기 정규세션

ToBig's 14기 강의자

강재영

Regression Analysis

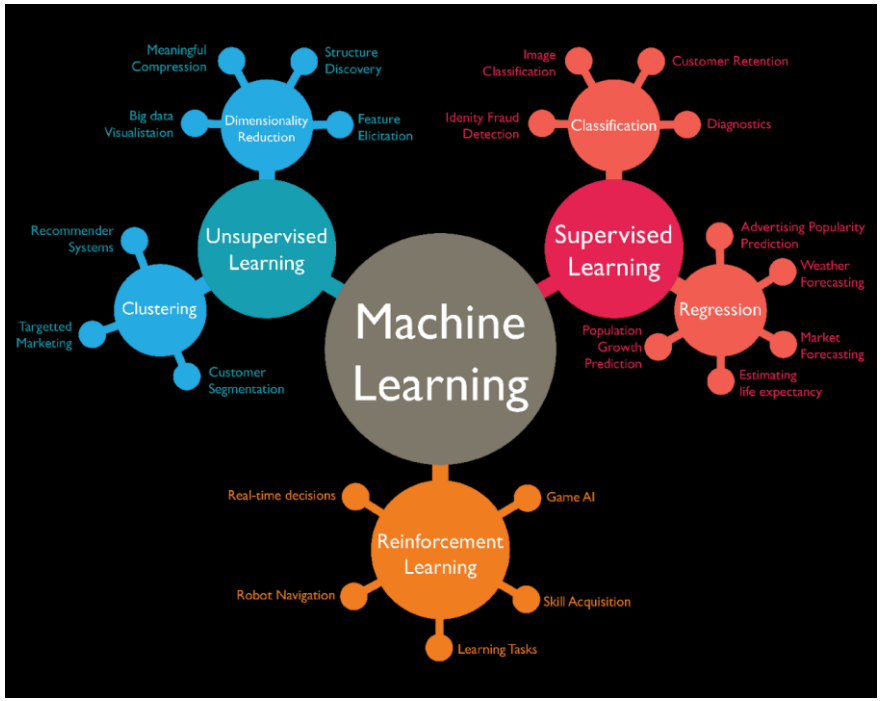
회귀분석

Contents

Unit 01 | 선형 회귀분석

머신러닝 알고리즘 종류

1. 지도학습 (Supervised Learning)
   - 입력과 결과값 (label : 정답) 이용한 학습
   - 회귀 (Regression), 분류 (Classification)
   - ex. **선형 / 로지스틱 회귀**, KNN, SVM, Decision Tree



2. 비지도학습 (Unsupervised Learning)
   - 입력만을 이용한 학습
   - 군집화 (Clustering)
   - ex. K-means Clustering

3. 강화학습 (Reinforcement Learning)
   - Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습

선형 회귀분석

- Input과 Output 사이의 선형관계를 도출하고자 하는 기법
- **Example)** 집의 가격(Output)과 집의 크기(Input)의 관계

Size (x)	Price (y)
2104	460
1416	232
1534	315
852	178
⋮	⋮

- **How?** Input과 Output 사이의 선형관계를 **어떻게** 알아낼 수 있을까?

*** 용어정리**
x (Input) : 영향을 주는 변수 – 독립변수, 설명변수
y (Output) : 영향을 받는 변수 – 종속변수, 반응변수

Unit 01 | 선형 회귀분석

Input (x)	Output (y)	Estimate ($\hat{y}$)
0	3	?
1	6	?
2	7	?
3	9	?

Parameter (모수)

Formulation : $h_\theta(x) = \theta_0 + \theta_1 x$

Hypothesis (가설)

e.g., $\theta_0 = 1,\ and\ \theta_1 = 2 \rightarrow h_\theta(x) = 1 + 2x$

*** 용어정리**

Parameter (모수) : 우리가 추정해야 할 수 - 회귀계수 $\theta_0\ \theta_1$

Y_hat $(\hat{y})$: 예측값

Unit 01 | 선형 회귀분석

Input (x)	Output (y)	Estimate ($\hat{y}$)
0	3	?
1	6	3
2	7	?
3	9	?

Parameter (모수)

Formulation : $h_\theta(x) = \theta_0 + \theta_1 x$

Hypothesis (가설)

e.g., $\theta_0 = 1,\ and\ \theta_1 = 2 \rightarrow h_\theta(x) = 1 + 2x$

1

Think about the case $x = 1$ (Loss = 6 − 3 = 3)

*** 용어정리**

Parameter (모수) : 우리가 추정해야 할 수 - 회귀계수 $\theta_0\ \theta_1$

Y_hat $(\hat{y})$: 예측값

Unit 01 | 선형 회귀분석

Input (x)	Output (y)	Estimate ($\hat{y}$)
0	3	?
1	6	3
2	7	?
3	9	?

Parameter (모수)

Formulation : $h_\theta(x) = \theta_0 + \theta_1 x$

Hypothesis (가설)

e.g., $\theta_0 = 1,\ and\ \theta_1 = 2 \rightarrow h_\theta(x) = 1 + 2x$

1

Think about the case $x = 1$ (Loss = 6 − 3 = 3)

▪**How?** Input과 Output 사이의 최적의 선형관계를 **어떻게** 알아낼 수 있을까?

*** 용어정리**

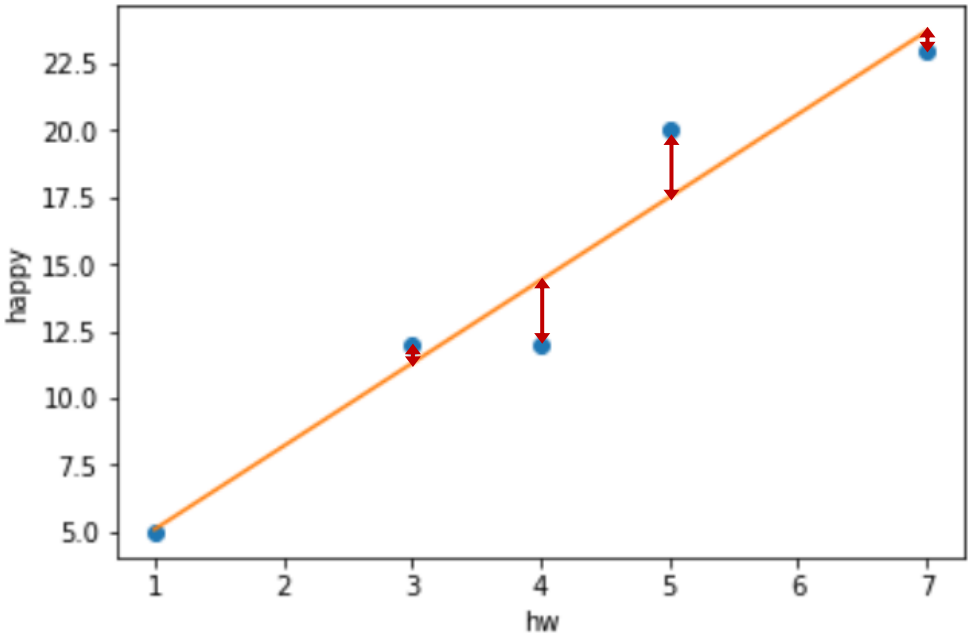
Parameter (모수) : 우리가 추정해야 할 수 - 회귀계수 $\theta_0\ \theta_1$

Y_hat ($\hat{y}$) : 예측값

Unit 01 | 선형 회귀분석

▪**How?** Input과 Output 사이의 최적의 선형관계를 **어떻게** 알아낼 수 있을까?

최소제곱법 (LSE)



회귀식이 예측한 값과 실제 값의 차이 최소화

회귀식이 예측한 $\hat{y}$ 값과 실제 y값의 차이의 제곱합을 최소화하는 알고리즘

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

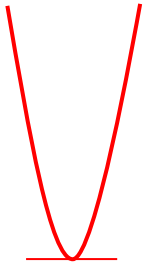
Loss Function

Unit 01 | 선형 회귀분석

최소제곱법 (LSE)

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Loss Function
목적함수



최소화하기위해 편미분 = 0

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

<정규방정식 : Normal Equation>

<최소제곱 추정치>

$$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Unit 01 | 선형 회귀분석

▪최소제곱법 (LSE)의 기하학적 관점

선형방정식을 행렬표현

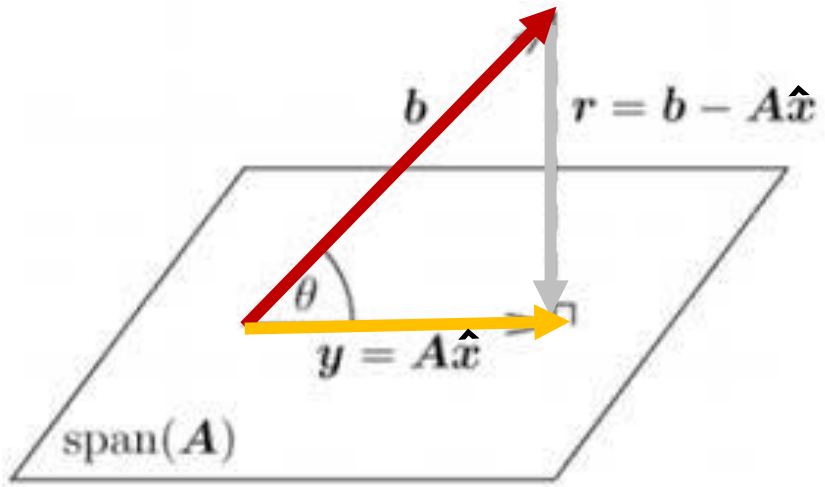
$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$\mathbf{Ax} = \mathbf{b}$$

Unit 01 | 선형 회귀분석

▪최소제곱법 (LSE)의 기하학적 관점



예측한 값과 실제 값의 차이 최소화

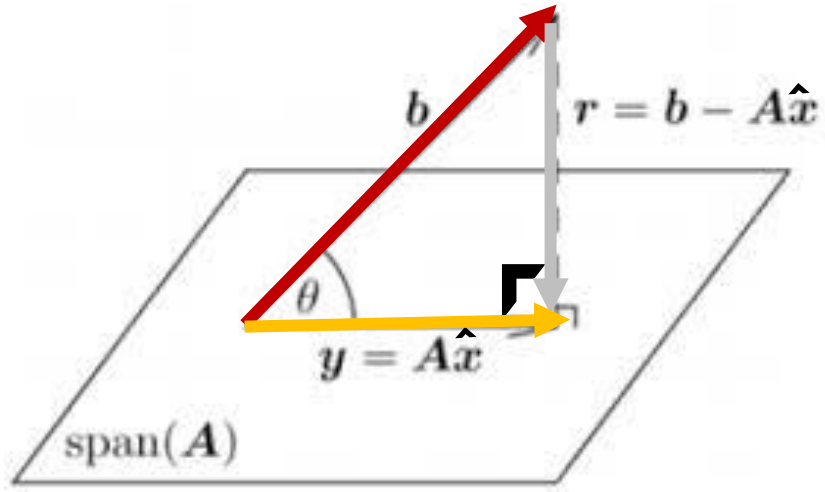
$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
&a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
&\vdots \\
&a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\end{aligned}
$$

$$
\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}
=
\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}
$$

$$\mathbf{Ax = b}$$

Unit 01 | 선형 회귀분석

▪최소제곱법 (LSE)의 기하학적 관점



< Projection >

예측한 값과 실제 값의 차이 최소화

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

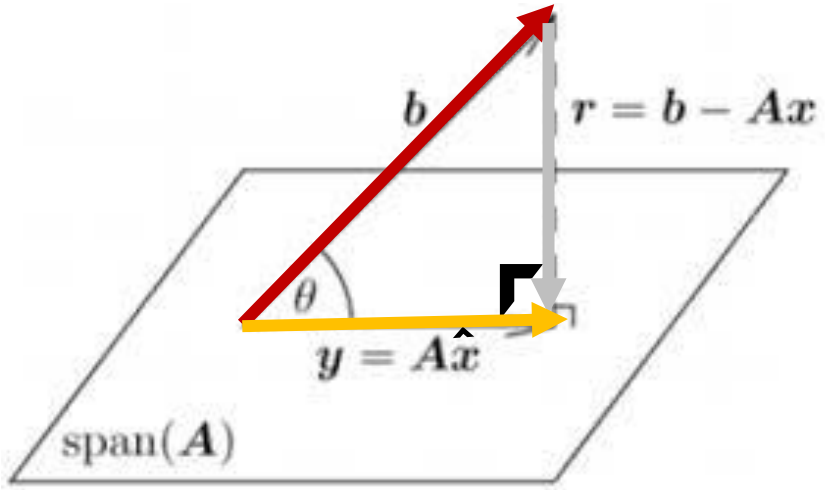
$$\mathbf{Ax} = \mathbf{b}$$

***용어정리**

Projection :한 벡터에서 다른차원의 공간으로 가장 최단거리가 되도록 선을 내리는 것

Unit 01 | 선형 회귀분석

▪최소제곱법 (LSE)의 기하학적 관점



< Projection >

선형방정식을 행렬표현

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$\mathbf{Ax = b}$$

다중회귀의 행렬표현

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \cdots & \cdots & & \cdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}$$

***용어정리**

Projection :한 벡터에서 다른차원의 공간으로 가장 최단거리가 되도록 선을 내리는 것

Unit 01 | 선형 회귀분석

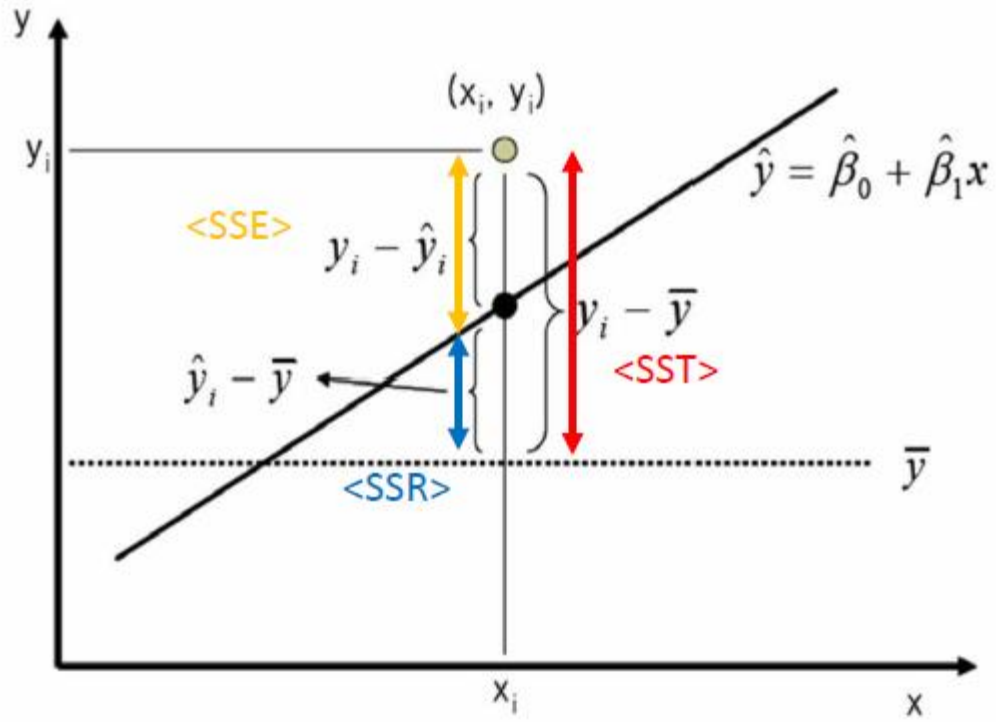
제곱합 분해

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

<SST> <SSR> <SSE>

SST : 총제곱합
SSR : 회귀제곱합 (전체 제곱합 중 회귀식으로 설명 Yes)
SSE : 잔차제곱합 (전체 제곱합 중 회귀식으로 설명 No)

- 회귀식이 데이터를 잘 설명할수록 SSR 증가 (SSE감소)



회귀분석 표 해석

	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		

$H0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$
F > F(α, p, n-p-1) 이면 H0 기각

MSE = 회귀식이 설명하지 못하는 부분
-> MSE 값이 작을수록 좋음

(cf) 단순회귀 제약조건 2개 = 1+1개
1. $\sum$ 잔차 = 0 2. $\sum$ 잔차 * x_i = 0

Contents

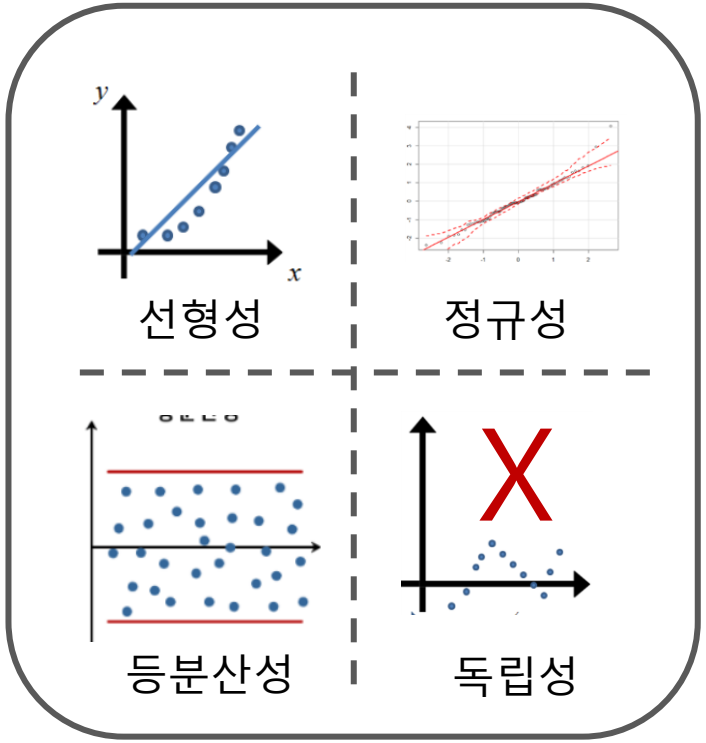
Unit 02 | 회귀 진단

회귀진단

완전하고 유용한 데이터 분석을 수행하기 위해 **기본가정들과 모형에 대한 문제점을 검출하고 수정**하는 것

회귀분석 기본 가정

1. 선형성 : 설명변수(X)와 반응변수(Y)가 선형 관계에 있다
2. 정규성 : $\varepsilon_i \sim N(0, \sigma^2)$, 오차(error) ε_i 는 정규분포를 따른다
3. 등분산성 : 오차 (error) ε_i 의 분산은 σ^2로 항상 동일하다
4. 독립성 : 오차 (error) ε_i 는 서로 독립이다 (iid)

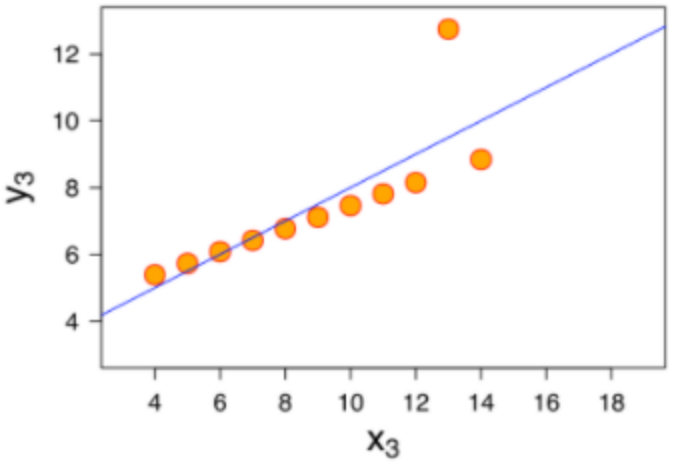


Unit 02 | 회귀 진단

그래프적 방법들

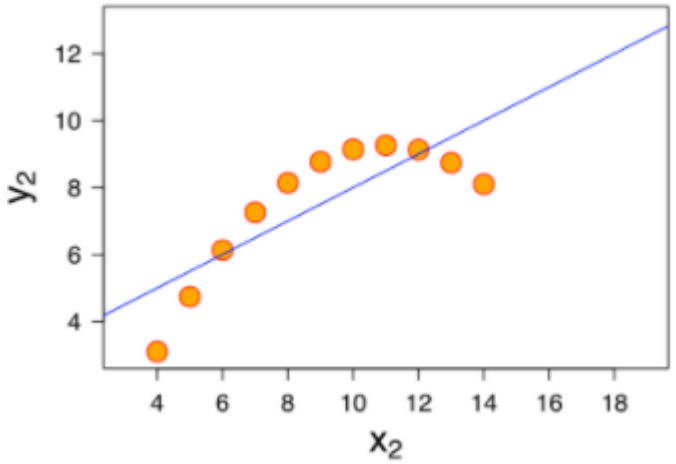
1) 선형성 판단

<Y와 X의 Scatter plot>



선형성 가정 O

<Y와 X의 Scatter plot>



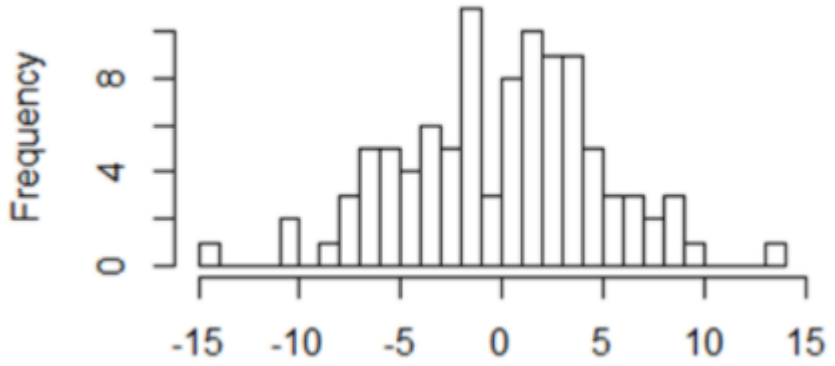
선형성 가정 X

Unit 02 | 회귀 진단

그래프적 방법들

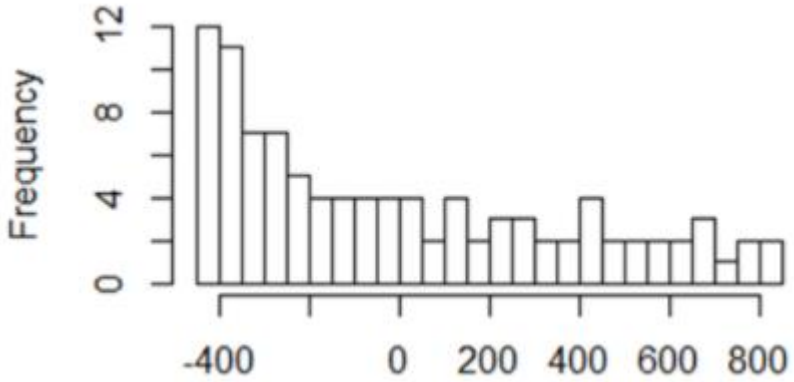
2) 정규성 판단

<잔차의 히스토그램>



정규성 가정 O

<잔차의 히스토그램>



정규성 가정 X

*** 용어정리**

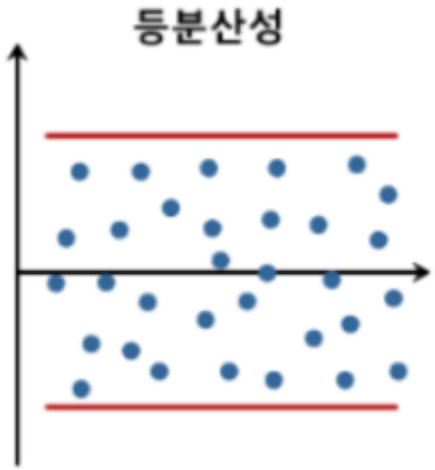
오차 (모수) = 잔차 (표본)

Unit 02 | 회귀 진단

그래프적 방법들

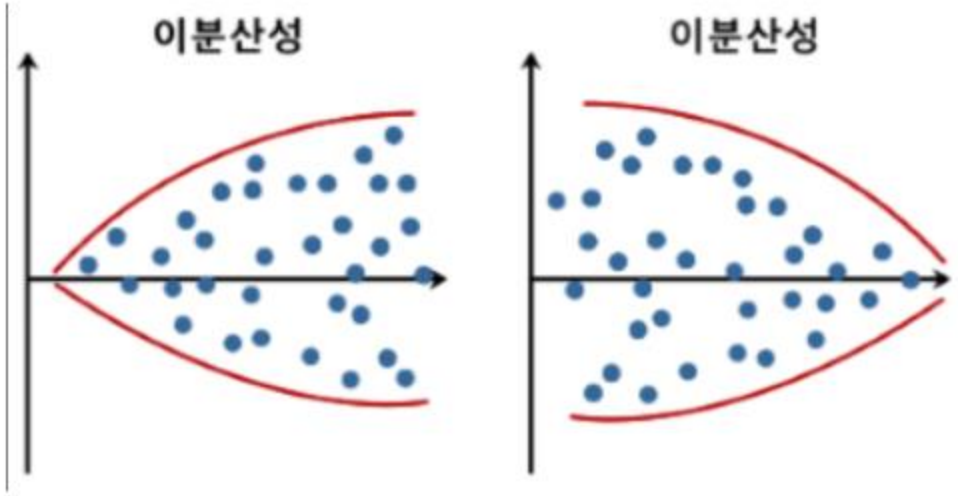
3) 등분산성 판단

<잔차의 산점도>



등분산성 O

<잔차의 산점도>



등분산성 X

Unit 02 | 회귀 진단

OLS Regression Results

Dep. Variable:	OPS	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	1931.
Date:	Tue, 28 Jul 2020	Prob (F-statistic):	0.00
Time:	02:03:49	Log-Likelihood:	254.44
No. Observations:	1633	AIC:	-490.9
Df Residuals:	1624	BIC:	-442.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.3380	0.018	18.524	0.000	0.302	0.374
BB	0.3019	0.059	5.151	0.000	0.187	0.417
HBP	0.1914	0.043	4.411	0.000	0.106	0.277
SO	0.0439	0.051	0.854	0.393	-0.057	0.145
height	0.2135	0.032	6.701	0.000	0.151	0.276
age_year	0.2850	0.024	11.762	0.000	0.237	0.333
HR	0.0194	0.009	2.064	0.039	0.001	0.038
SB	0.0052	0.007	0.749	0.454	-0.008	0.019
H	0.0293	0.013	2.217	0.027	0.003	0.055

Omnibus:	580.341	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8336.581
Skew:	1.255	Prob(JB):	0.00
Kurtosis:	13.780	Cond. No.	17.3

OLS : ordinary least square

- R-squared / Adj. R-squared
- F-statistics
- Coefficients p값
- Durbin-Watson (오차의 자기상관)

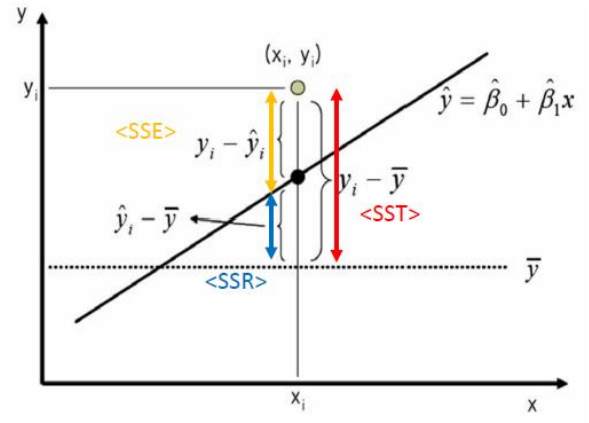
Unit 02 | 회귀 진단

모형 선택 기준 : R-squared(결정계수)

```
R-squared:                       0.915
Adj. R-squared:                  0.914
```

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

전체 제곱합 중 회귀식으로 설명 가능한 부분
-> 결정계수가 크면 클수록 좋음 !



SST : 총제곱합
SSR : 회귀제곱합
SSE : 잔차제곱합

$$adj\ R^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Adjusted R square(조정된 결정계수)

설명변수를 추가하면 SSR 이 항상 커져 결정계수가 증가
따라서 설명변수의 개수가 다른 모델과 비교 어려움
-> 설명변수의 개수를 고려하여,
설명변수가 증가하면 값이 감소하도록 패널티를 줌

Unit 02 | 회귀 진단

교호작용 (interaction)

- 단일 변수만으로는 알 수 없는 변수들간의 상호작용 고려

- 보통 범주형*범주형 , 범주형*연속형 변수의 관계만 고려
  -> 연속형*연속형의 경우 해석의 모호함이 생길 수 있기 때문에 !

- ex) 흡연을 하면 건강 -3, 음주를 하면 건강 -2 ⇔ 흡연과 음주를 동시에 하는 사람은 ?
  흡연과 음주를 동시에 한 효과 더 많은 악영향 (-10)이 끼친다면, 교호작용 변수를 고려해야 함 !
- 건강(Y) ~ 흡연 + 음주 → 건강(Y) ~ 흡연 + 음주 + 흡연*음주

Unit 02 | 회귀 진단

https://m.blog.naver.com/vnf3751/220830413960

* p-value? 귀무가설? 유의확률 ?

p값이 0.05보다 작으므로
95% 유의수준 하에서 귀무가설을 기각한다

가설검정
모집단의 특징에 대한 통계적 가설을
추출된 표본을 통하여 검토하는 추론 방법

귀무가설 (H0) : 기각하고자 하는 사실
대립가설 (H1) : 일반적으로 주장하고자 하는 사실

> H0를 기각함으로써, H1을 입증한다 !

$$H_0: \beta_1 = 0$$

$$\text{성적} = \beta_0 + \beta_1 \text{지능}$$

$$H_1: \beta_1 \neq 0$$

만약 여러분이 성적과 지능은 유의미한 상관관계가 있다고 주장한다면, 가설은 다음과 같이 세워질 수 있습니다.

Unit 02 | 회귀 진단

https://m.blog.naver.com/vnf3751/220830413960

* p-value? 귀무가설? 유의확률 ?

$$\text{성적} = \beta_0 + \beta_1 \text{지능}$$

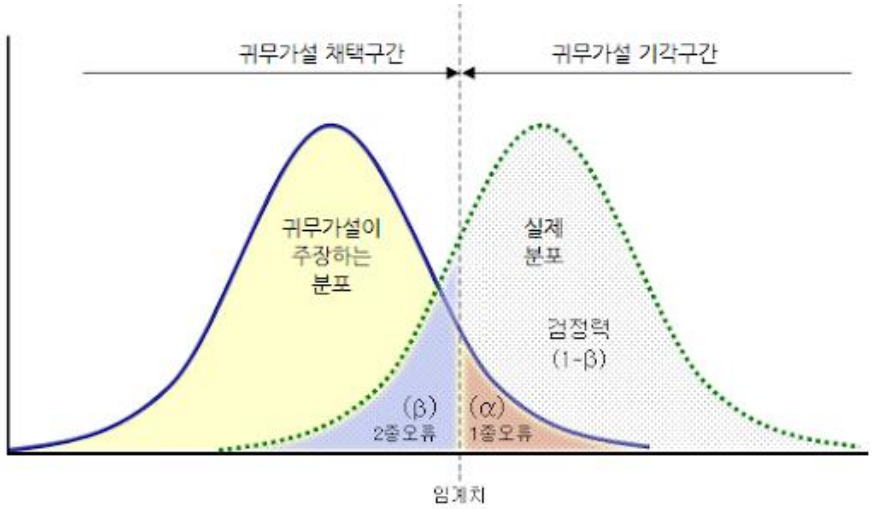
$$y = \beta_0 + \beta_1 x$$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

검정통계량 : 귀무가설이 옳다는 가정 하에서 구하게 된 통계량

P-value : 귀무가설이 옳다는 가정 하에, 검정통계량이 계산될 확률



- 제1종오류 : 귀무가설이 참인데 기각하는 경우
- 제2종오류 : 귀무가설이 거짓인데 채택하는 경우

- 귀무가설이 옳은데 실수로 기각될 확률, 즉, 1종 오류를 범하게 될 확률 최소화
- 1종 오류의 상한선 (=유의수준) 미리 설정 (일반적으로 0.05)

Unit 02 | 회귀 진단

F – Statistics & t- Statistics

OLS Regression Results

Dep. Variable:	OPS	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	1931.
Date:	Tue, 28 Jul 2020	Prob (F-statistic):	0.00
Time:	02:03:49	Log-Likelihood:	254.44

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0, \quad for\ some\ j$$

F- 통계량은 모형의 유의미함을 판단하는 기준으로 모든 독립 변수의 계수가 0인지 혹은 하나라도 0이 아닌지를 판별합니다.

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.3380	0.018	18.524	0.000	0.302	0.374
BB	0.3019	0.059	5.151	0.000	0.187	0.417

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

t-통계량은 변수의 유의미함을 판단하는 기준으로 해당 변수의 계수가 0인지 아닌지를 판별합니다.

Durbin-Watson (오차의 자기상관)

OLS Regression Results

Dep. Variable:	OPS	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	1931.
Date:	Tue, 28 Jul 2020	Prob (F-statistic):	0.00
Time:	02:03:49	Log-Likelihood:	254.44
No. Observations:	1633	AIC:	-490.9
Df Residuals:	1624	BIC:	-442.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.3380	0.018	18.524	0.000	0.302	0.374
BB	0.3019	0.059	5.151	0.000	0.187	0.417
HBP	0.1914	0.043	4.411	0.000	0.106	0.277
SO	0.0439	0.051	0.854	0.393	-0.057	0.145
height	0.2135	0.032	6.701	0.000	0.151	0.276
age_year	0.2850	0.024	11.762	0.000	0.237	0.333
HR	0.0194	0.009	2.064	0.039	0.001	0.038
SB	0.0052	0.007	0.749	0.454	-0.008	0.019
H	0.0293	0.013	2.217	0.027	0.003	0.055

Omnibus:	580.341	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8336.581
Skew:	1.255	Prob(JB):	0.00
Kurtosis:	13.780	Cond. No.	17.3

더빈 왓슨(Durbin Watson) 검정

: **오차항이 독립성을 만족하는지를 검정하기 위해 사용**

- 더빈 왓슨 통계량은 0 ~ 4사이의 값을 갖을 수 있음

  0에 가까울수록 → 양의 상관관계

  4에 가까울수록 → 음의 상관관계

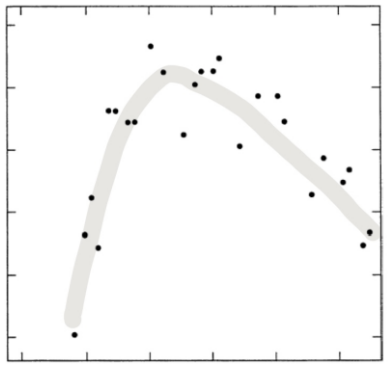
  2에 가까울수록 → 오차항의 자기상관이 없음 (독립성만족)

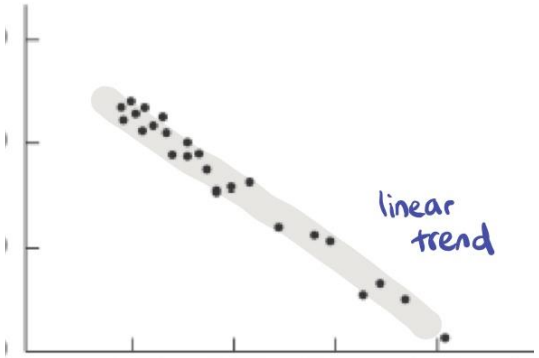
*자기상관(Autocorrelation) : 오차항이 서로 상관관계가 존재하는 경우

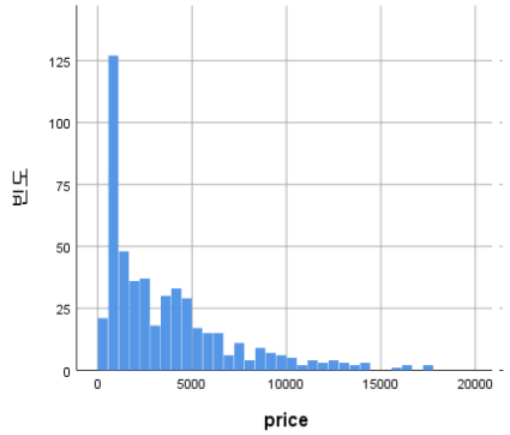
Unit 02 | 회귀 진단

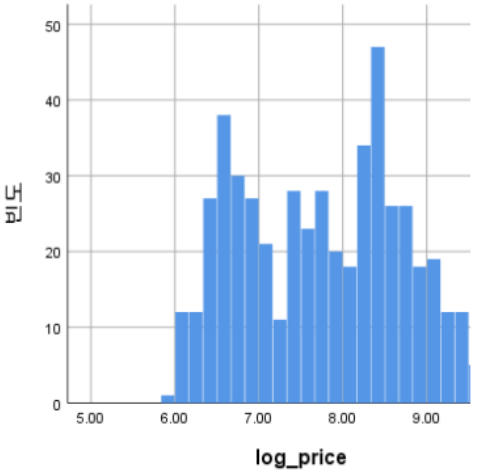
변수 변환

https://every-day-life.tistory.com/16









1) 비선형적인 함수 관계를 선형으로 바꿔 다룰 수 있다
2) 분포모양을 정규분포와 유사하도록 만들 수 있다
3) 변환을 통해 자기상관 문제를 해결 할 수 있다

ex) $\log(x)$, $\sqrt{x}, x^2$, ...

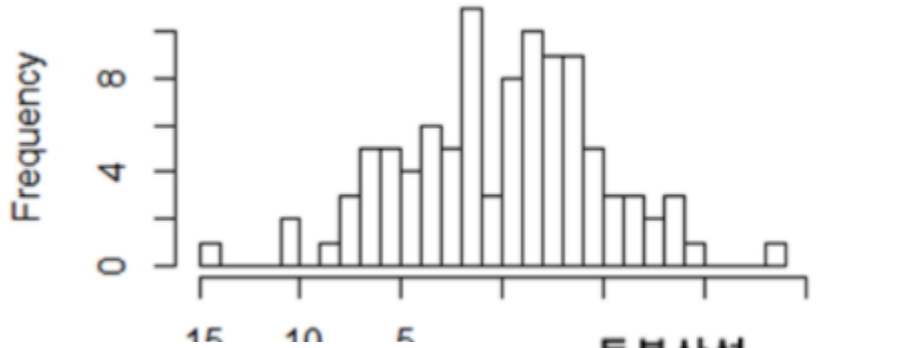
Unit 02 | 회귀 진단

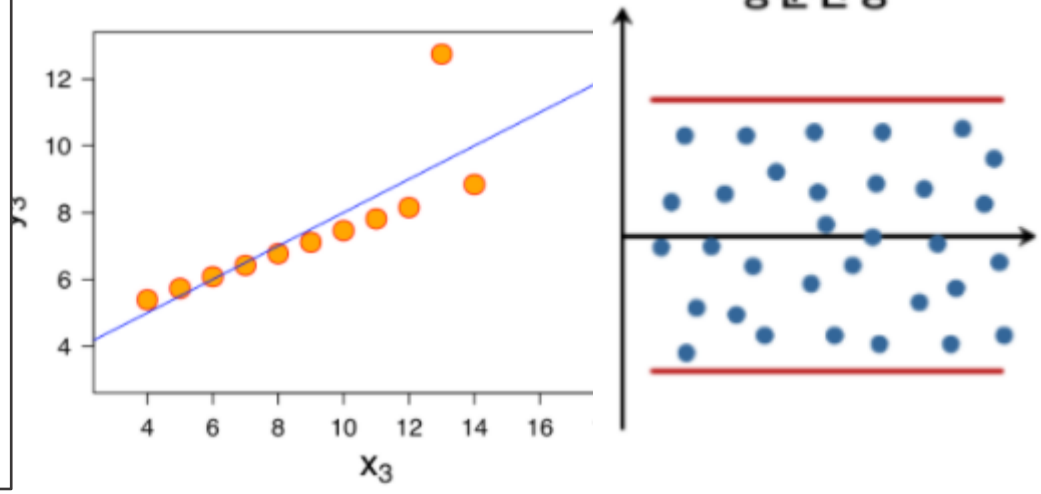
OLS Regression Results

Dep. Variable:	OPS	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	1931.
Date:	Tue, 28 Jul 2020	Prob (F-statistic):	0.00
Time:	02:03:49	Log-Likelihood:	254.44
No. Observations:	1633	AIC:	-490.9
Df Residuals:	1624	BIC:	-442.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.3380	0.018	18.524	0.000	0.302	0.374
BB	0.3019	0.059	5.151	0.000	0.187	0.417
HBP	0.1914	0.043	4.411	0.000	0.106	0.277
SO	0.0439	0.051	0.854	0.393	-0.057	0.145
height	0.2135	0.032	6.701	0.000	0.151	0.276
age_year	0.2850	0.024	11.762	0.000	0.237	0.333
HR	0.0194	0.009	2.064	0.039	0.001	0.038
SB	0.0052	0.007	0.749	0.454	-0.008	0.019
H	0.0293	0.013	2.217	0.027	0.003	0.055

Omnibus:	580.341	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8336.581
Skew:	1.255	Prob(JB):	0.00
Kurtosis:	13.780	Cond. No.	17.3



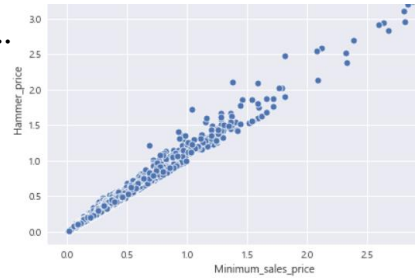


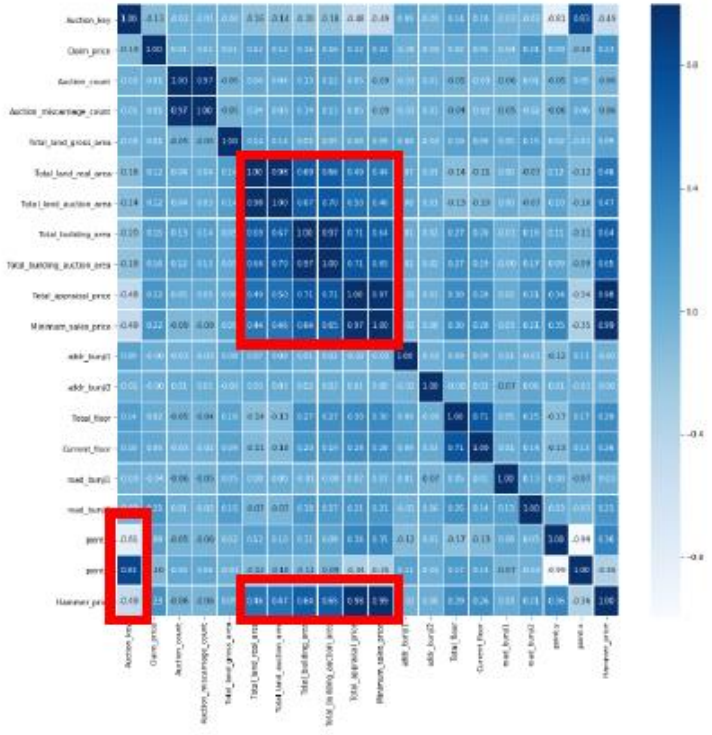
Unit 02 | 회귀 진단

다중공선성 (Multicollinearity)

- 설명변수 간 상관관계가 강해 설명변수의 일부를 다른 설명변수의 선형결합으로 표현가능한 것

ex) X2=2X1, X3=X1+X2, ..





- 상관계수, scatter plot, heatmap 등 확인

- 다중공선성 진단통계량 : VIF (Variance inflation factor)
  - i 번째 설명변수를 다른 설명변수들로 회귀한 성능
  - 다른 설명변수들과 상관관계가 강할수록 VIF 값이 큼
  - VIF > 10 이면 다중공선성 존재한다고 판단 !

$$\mathrm{VIF}_i = \frac{\sigma^2}{(n-1)\mathrm{Var}[X_i]} \cdot \frac{1}{1-R_i^2}$$

Unit 02 | 회귀 진단

다중공선성을 제거하는 이유 ?

- 설명변수 간 독립적이지 않으면 회귀계수의 추정이 불안정하게 됨 !
- 추정값이 존재하지 않거나, 추정값의 분산이 매우 매우 커지거나 …

$$\hat{\beta} = (X'X)^{-1}X'y$$

설명변수끼리 완벽한 선형관계가 존재하면
이 부분이 Full rank가 아니어서 역행렬 존재하지 않음

완벽한 선형관계가 아니더라도, 강한 다중공선성이 존재하면
이 부분이 작아서 역행렬을 취하면 값이 매우 커짐

-> 회귀 계수의 분산이 매우 커지게 되어 불안정한 추정이 됨

Ex) 학업성취도(Y) ~ 일평균 음주량(X1), 일평균 혈중 알코올농도(X2)

B1_hat = -0.5 , B2_hat = - 24

다중공선성 제거 방법

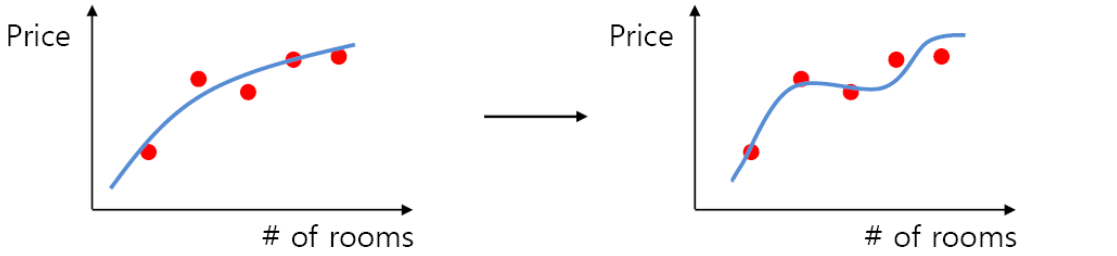
1. 더 많은 데이터 수집
2. 상관관계 가장 높은 변수 제거
3. PCA : 차원 축소 (dimension reduction) -> 향후 강의 있을 예정 ..
4. Ridge / Lasso Regression

과적합(Overfitting)

- 모형이 Train data에만 너무 딱 맞게 적합되어서, 실제 data에는 성능이 낮게 나오는 경우

정규화 (Regularization)

- 모델이 복잡해질수록 penalty를 크게 주도록, 목적 함수에 항을 하나 더 추가



$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$\min_\theta J(\theta), \text{ where } J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + 1000\theta_3^2 + 1000\theta_4^2$$

Loss Function

Unit 02 | 회귀 진단

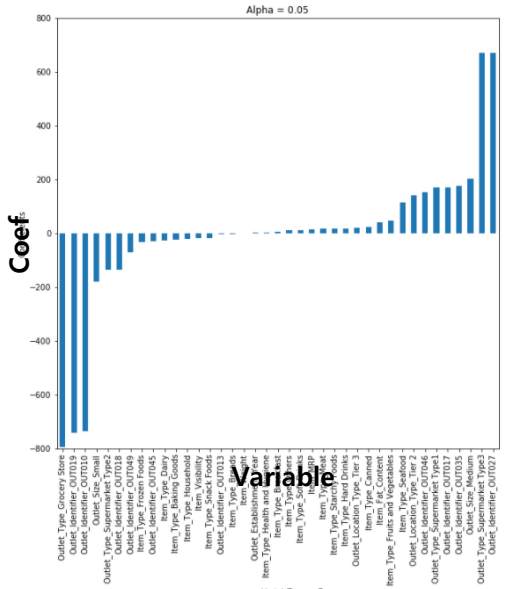
정규화 (Regularization)

1. Ridge Regression

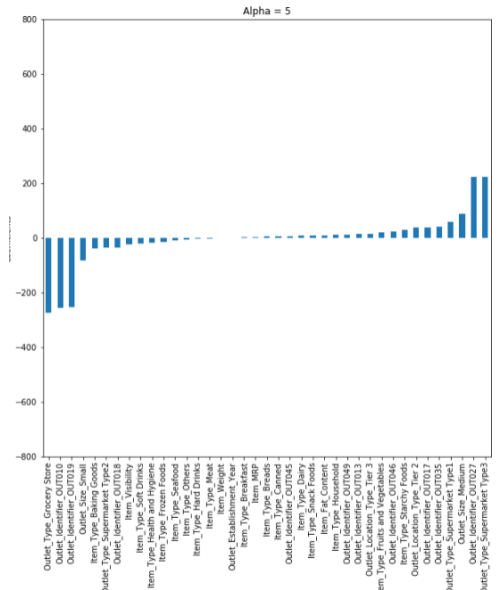
$$J(\theta) = \text{MSE}(\theta) + \alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

Loss Function

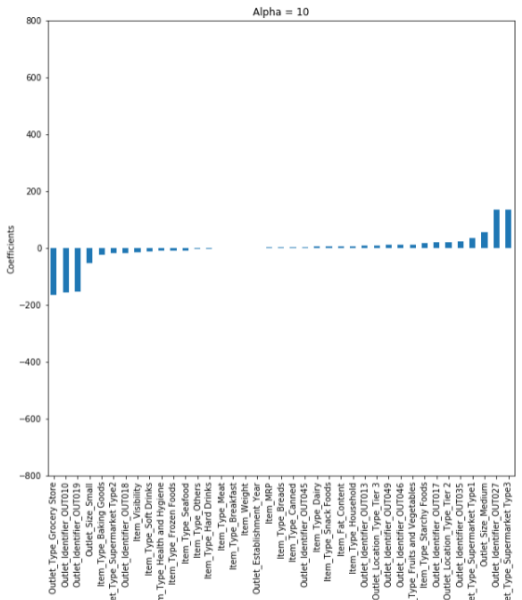
(α : 정규화 계수)



(α = 0.05)



(α = 5)



(α = 5)

Unit 02 | 회귀 진단

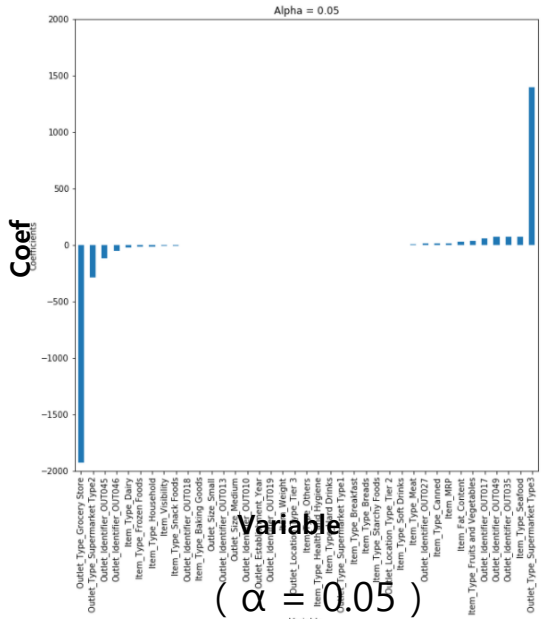
정규화 (Regularization)

2. Lasso Regression

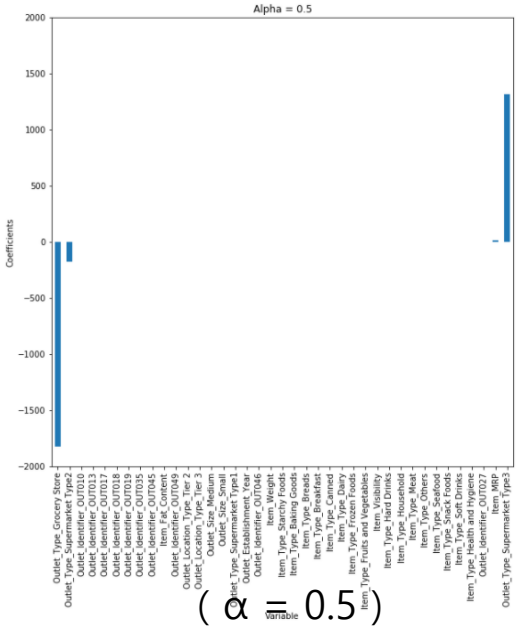
$$J(\theta) = \mathrm{MSE}(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

Loss Function

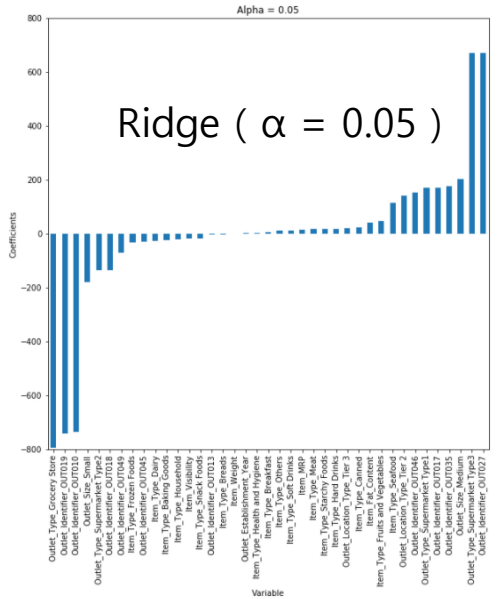
(α : 정규화 계수)



(α = 0.05)



(α = 0.5)



Unit 02 | 회귀 진단

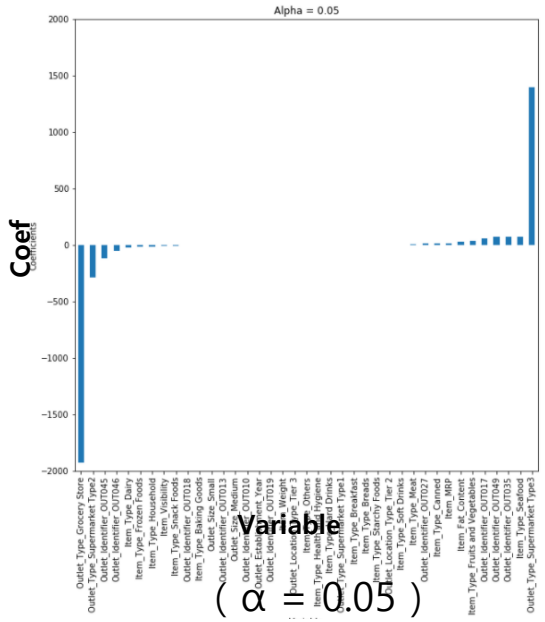
정규화 (Regularization)

2. Lasso Regression

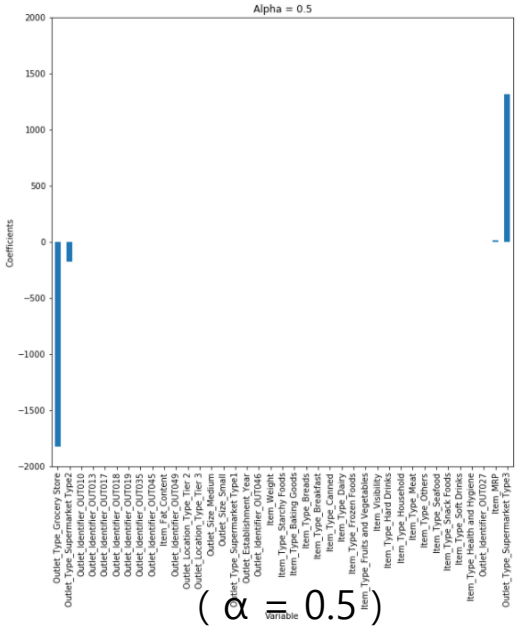
$$J(\theta) = \mathrm{MSE}(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

Loss Function

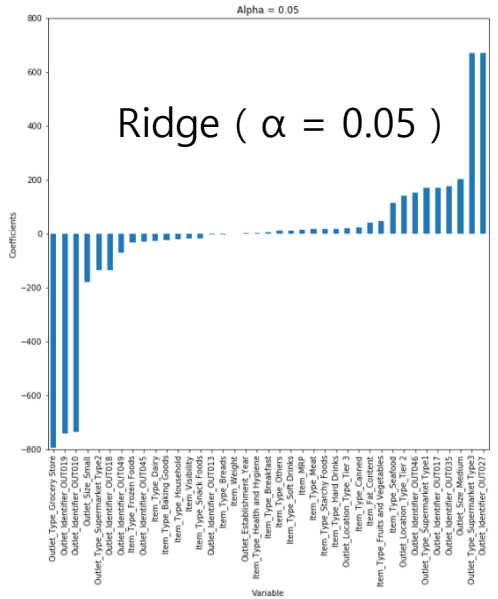
(α : 정규화 계수)



(α = 0.05)



(α = 0.5)



(+Elastic Net Regression)

https://brunch.co.kr/@itschloe1/11

Unit 02 | 회귀 진단

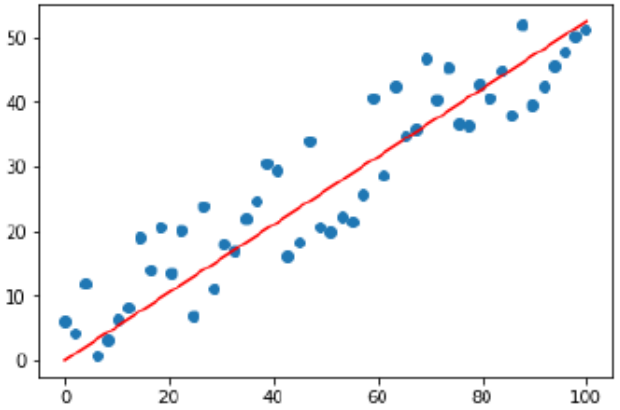
< 마무리 > 선형회귀분석

1. 회귀 모형 설정 : 반응변수 및 주요 설명변수 파악
2. 선형성 검토 : 산점도를 통해 상관관계 파악
3. 설명변수 검토 : 각 변수들의 분포 확인 + 다중공선성 파악
4. 모델 적합 : 모형의 회귀계수 추정 및 모형의 적절성 검토
5. 변수 선택 : 중요 설명변수 선택
6. 적합된 모형 검토 : 오차 가정 체크
7. 최종 모형 선택

Contents

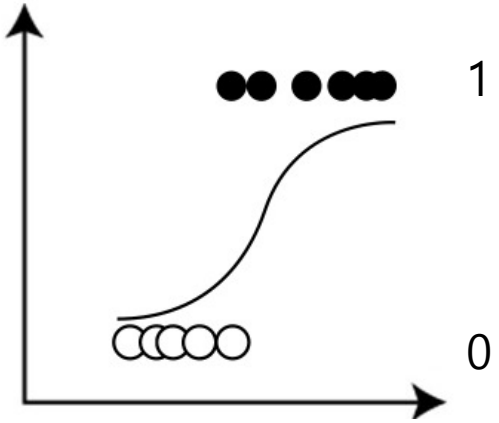
Unit 03 | 로지스틱 회귀분석

Linear Regression



연속형 Y 예측
Regression

Logistic Regression



범주형 Y 분류
Classification

로지스틱 회귀분석

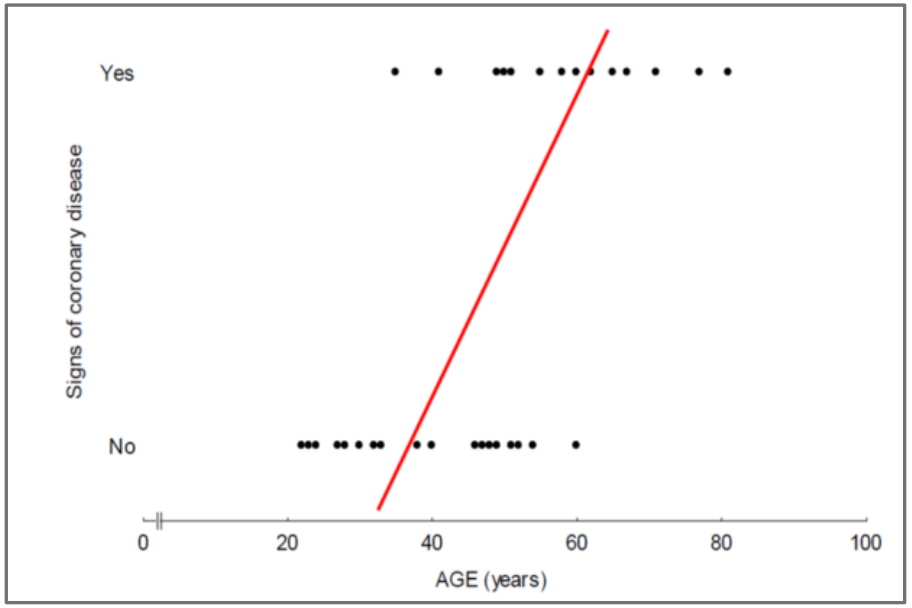
새로운 관측치가 있을 때, 이를 기존 범주 중 하나로 예측 (범주 예측)

로지스틱 회귀모델의 예시 : "분류"

- 제품이 불량인지 정상인지
- 고객이 이탈고객인지 잔류고객인지
- 카드 거래가 정상인지 사기인지
- 내원 고객이 질병이 있는지 없는지

Unit 03 | 로지스틱 회귀분석

범주형 변수를 선형회귀로 예측한다면 ...?



범위가 일치하지 않음 !

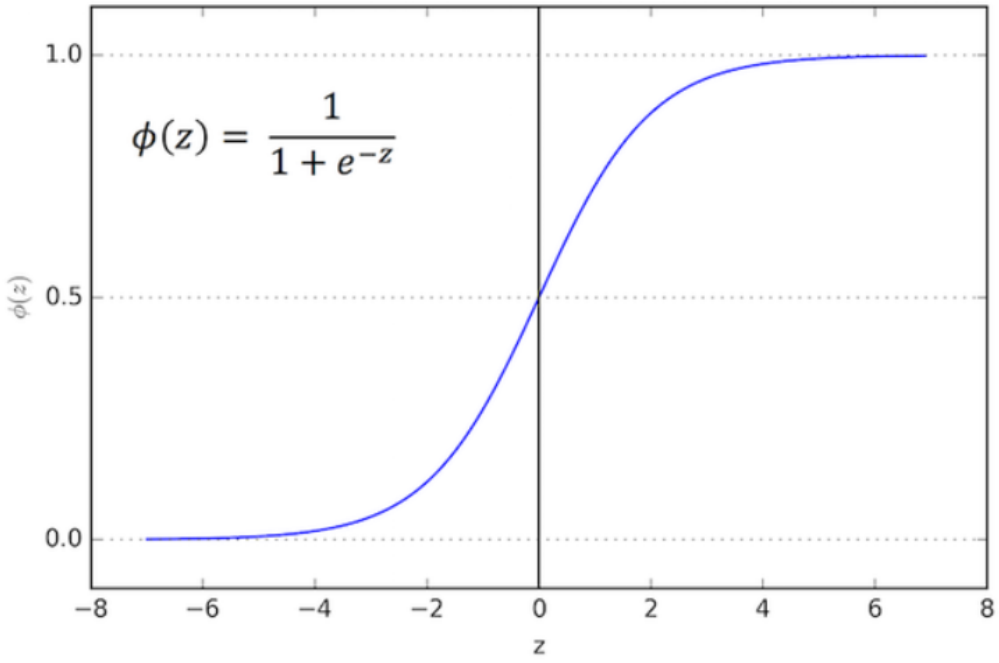
1. 선형회귀 (-inf, +inf)
2. 로지스틱 0 / 1

중간 범주가 없고, 숫자가 아무런 의미를 지니지 않게 됨

-> Y가 범주형(categorical) 변수일 때는
다중선형회귀 모델을 그대로 적용할 수 없다 !

Unit 03 | 로지스틱 회귀분석

Logistic function (Sigmoid function)



Output 범위 : (0, 1)
Input 값에 대해 단조증가 (or 단조감소)

$$f(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Unit 03 | 로지스틱 회귀분석

Odds(승산)

$$E(y) = \pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

β_1의 해석 직관적이지 못함 !

$$Odds = \frac{p}{1-p} = \frac{\text{성공확률}}{\text{실패확률}}$$

Odds (승산)

성공 확률을 p로 정의할 때,
실패 대비 성공 확률의 비율

로짓 변환 (Logit Transformation)

Logit function

$$\log(Odds) = \log\left(\frac{\pi(X=x)}{1-\pi(X=x)}\right) = \log\left(\frac{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}{1-\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}\right) = \beta_0 + \beta_1 x$$

$P(Y_i = 1) = \pi_i$
$P(Y_i = 0) = 1 - \pi_i$

β_1 의 의미 : x가 한 단위 증가했을 때, log(Odds)의 증가량

Odds Ratio(승산비)

$$\widehat{O_R} = \frac{Odds_{x+1}}{Odds_x} = e^{\widehat{\beta_1}}$$

OR > 1 : 독립변수가 종속변수에 양의 방향으로 영향을 미침
OR < 1 : 독립변수가 종속변수에 음의 방향으로 영향을 미침
OR = 1 : 독립변수가 종속변수에 영향을 미치지 않음

Unit 03 | 로지스틱 회귀분석

대출여부(0 or 1)에 관한 로지스틱 회귀분석

Logit Regression Results

Dep. Variable:	Personal Loan	No. Observations:	1750
Model:	Logit	Df Residuals:	1738
Method:	MLE	Df Model:	11
Date:	Fri, 23 Aug 2019	Pseudo R-squ.:	0.6030
Time:	14:55:31	Log-Likelihood:	-229.35
converged:	True	LL-Null:	-577.63
		LLR p-value:	2.927e-142

	coef	std err	z	P>|z|	[0.025	0.975]
Age	0.0245	0.102	0.240	0.810	-0.175	0.224
CCAvg	0.0985	0.063	1.562	0.118	-0.025	0.222
CD Account	4.3726	0.568	7.703	0.000	3.260	5.485
CreditCard	-1.2374	0.337	-3.667	0.000	-1.899	-0.576
Education	1.5203	0.190	7.999	0.000	1.148	1.893
Experience	-0.0070	0.102	-0.069	0.945	-0.206	0.192
Family	0.7579	0.128	5.914	0.000	0.507	1.009
Income	0.0547	0.004	12.659	0.000	0.046	0.063
Mortgage	-0.0001	0.001	-0.144	0.885	-0.002	0.002
Online	-0.4407	0.263	-1.674	0.094	-0.957	0.075
Securities Account	-1.8520	0.561	-3.299	0.001	-2.952	-0.752
const	-13.9203	2.773	-5.021	0.000	-19.354	-8.486

Age의 Coefficient = 0.0245

Age의 Coefficient = $\ln\left(\frac{Odds_{x+1}(\text{나이가 한살 더})}{Odds_{x}(\text{나이 그대로})}\right)$

따라서, e 변환으로 ln 제거 -> 오즈비만 남게됨

$e^{0.0245} = 1.024$,

다른 효과 동일, Age가 한 단위 증가할 때 대출(Y=1)할 확률이 1.024배 증가한다.

다른 효과 동일, Age가 한 단위 증가할 때 대출(Y=1)할 확률이 2.4% 증가한다.

회귀 계수의 해석

- Linear Regression : 설명변수가 1만큼 증가함에 따른 반응변수의 변화량

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$$

- Logistic Regression : 설명변수가 1만큼 증가함에 따른 로그 오즈의 변화량

$$\log(Odds) = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$$

Contents

Unit 04 | 최대우도추정

최적의 가설 (Optimal Hypothesis)

▪**How?** Input과 Output 사이의 최적의 가설 **어떻게** 알아낼 수 있을까?

- Linear Regression : $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$

  LSE(최소제곱법) $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

- Logistic Regression : $f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$

▪**How?** Input과 Output 사이의 최적의 가설 **어떻게** 알아낼 수 있을까?

LSE(최소제곱법) ???

Unit 04 | 최대우도추정

최적의 가설 (Optimal Hypothesis)

▪**How?** Input과 Output 사이의 최적의 가설 **어떻게** 알아낼 수 있을까?

- Linear Regression : $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$

  LSE(최소제곱법) $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

- Logistic Regression : $f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$

▪**How?** Input과 Output 사이의 최적의 가설 **어떻게** 알아낼 수 있을까? LSE(최소제곱법) (X)

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{\theta^T x^{(i)}}} - y^{(i)} \right)^2$$

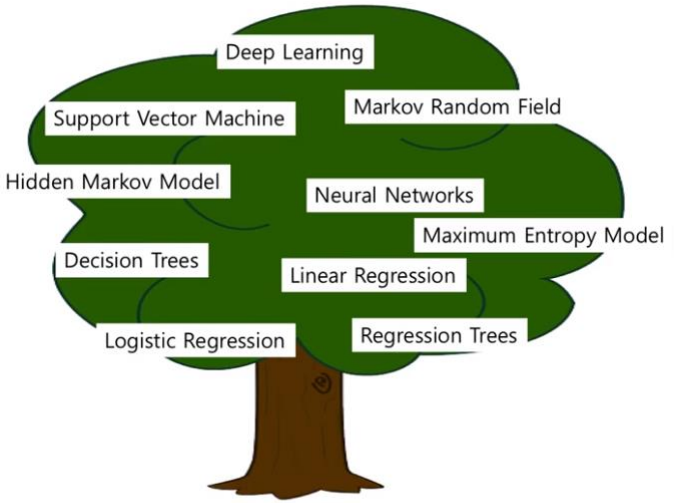
Not quadratic ! (i.e., non-convex)

Unit 04 | 최대우도추정

MLE (Maximum Likelihood Estimtion) : 최대 우도 추정법

선형회귀분석(최소제곱법)과 달리, MLE로 계수를 추정한다 !

$$\pi(X) = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_k} x_k)}}$$



최대 우도 추정 (Maximum Likelihood Estimation)

$$\hat{\theta} = \underset{\theta}{argmax}\, L(\theta)$$

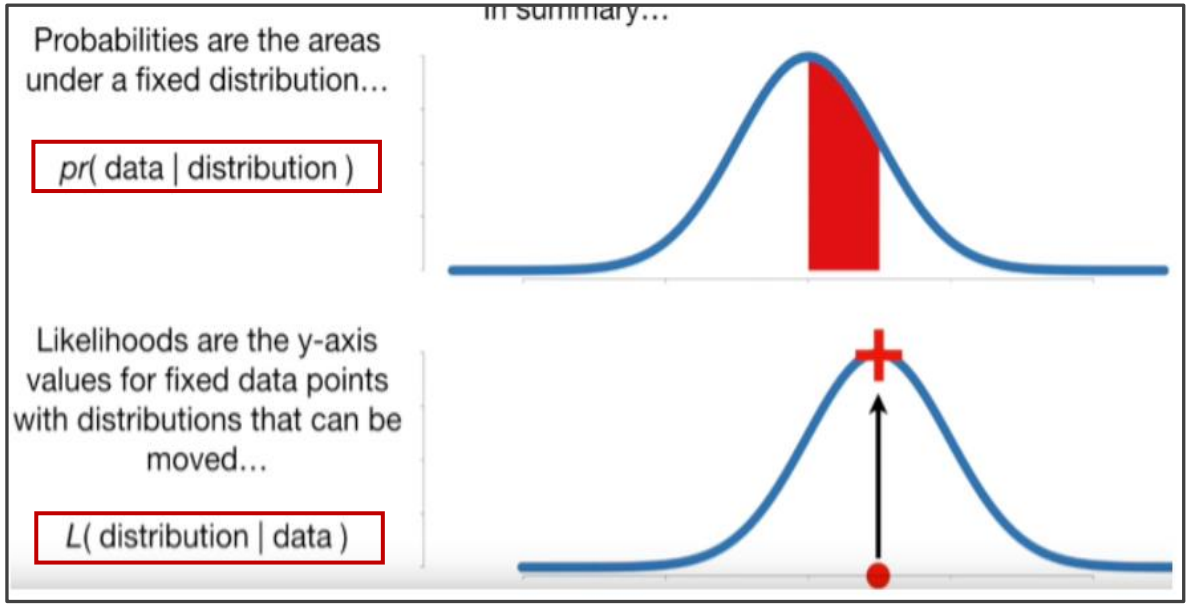
우도를 최대화하는 parameter 추정 !

Unit 04 | 최대우도추정

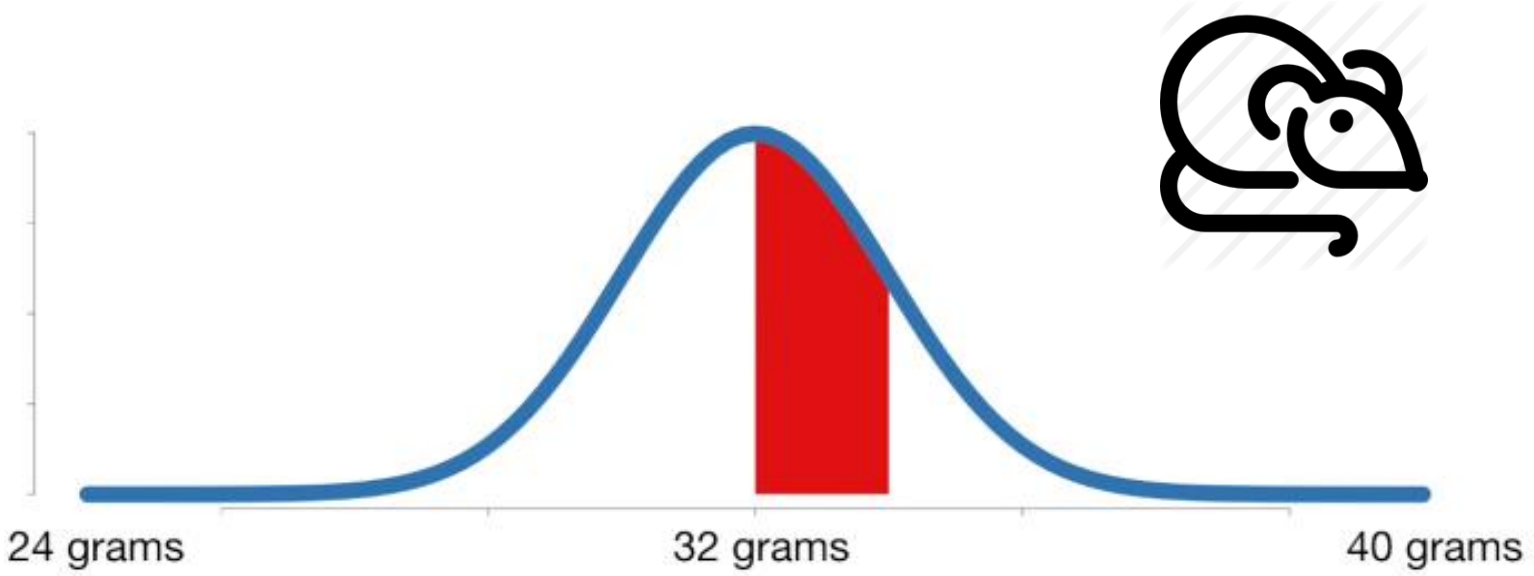
Likelihood(우도, 가능도)

- Probability
  주어진 확률분포에서 해당 관측값이 나올 확률

- Likelihood
  어떤 값이 관측되었을 때, 이것이 어떤 확률분포에서 왔을지에 대한 가능성



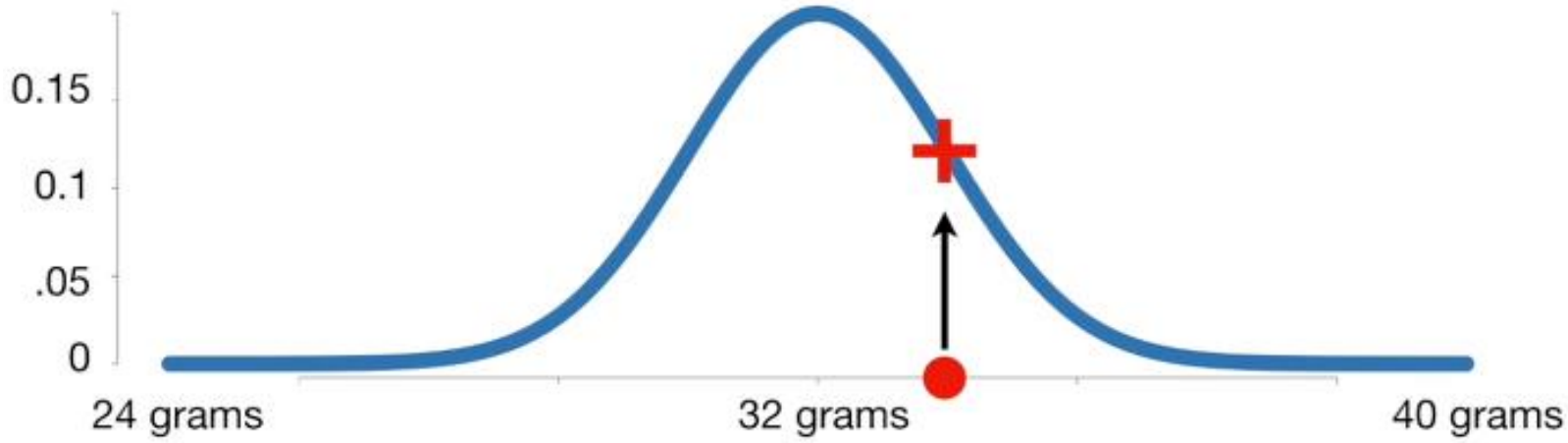
확률(Probability)



pr(weight between 32 and 34 grams | mean = 32 and standard deviation = 2.5)

Unit 04 | 최대우도추정

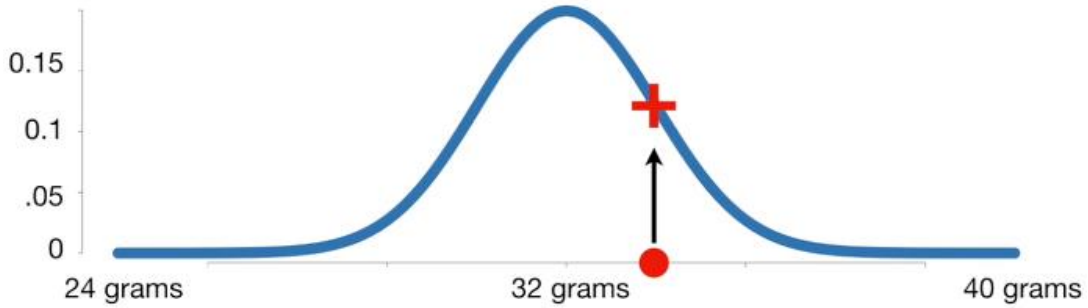
가능도(Likelihood)



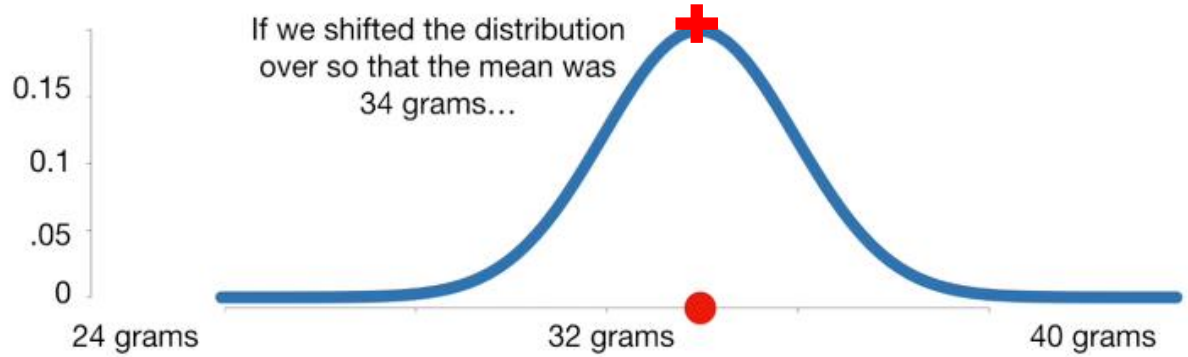
L(mean = 32 and standard deviation = 2.5 | mouse weighs 34 grams)

Unit 04 | 최대우도추정

가능도(Likelihood)



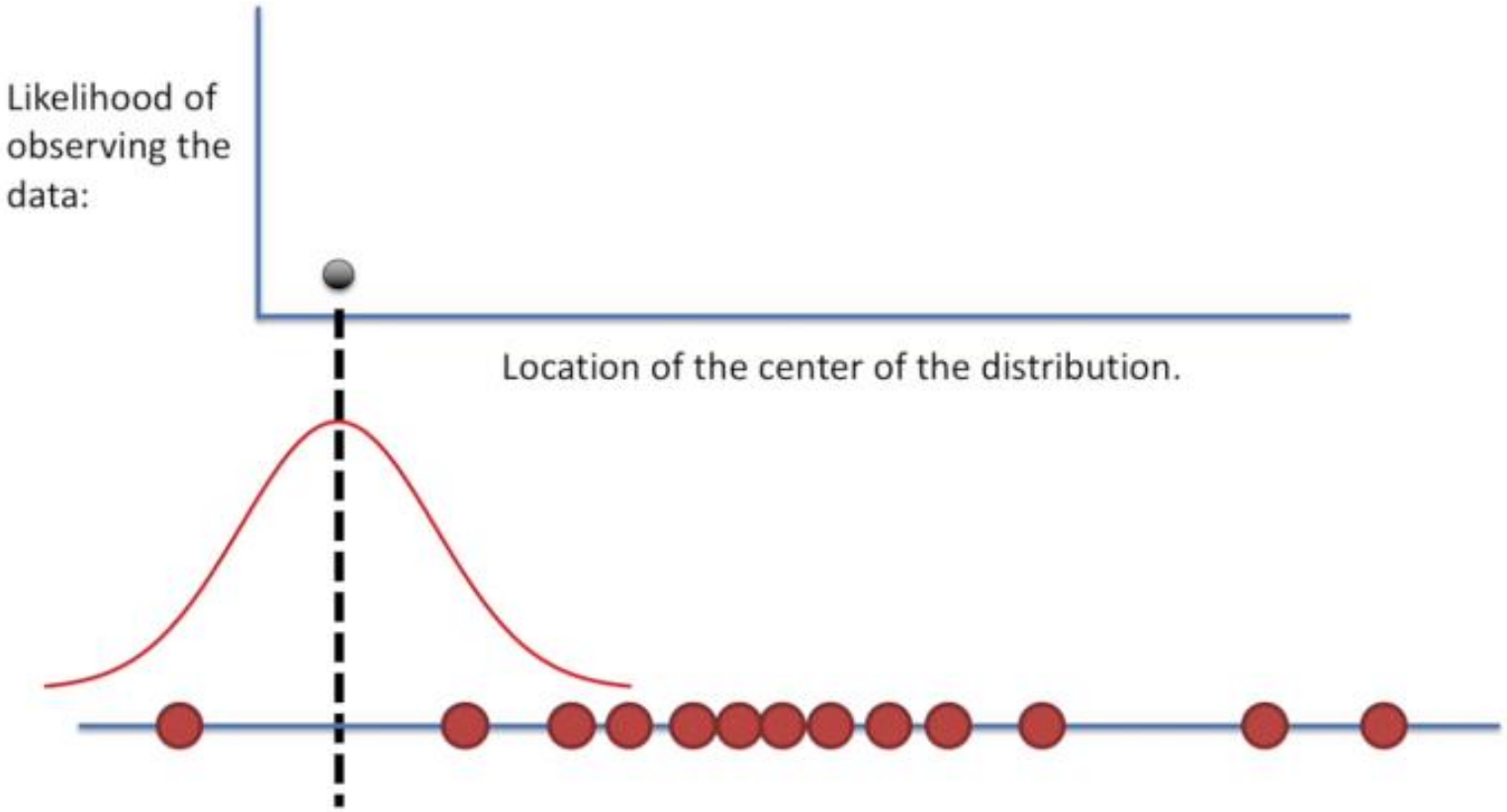
L(mean = 32 and standard deviation = 2.5 | mouse weighs 34 grams)



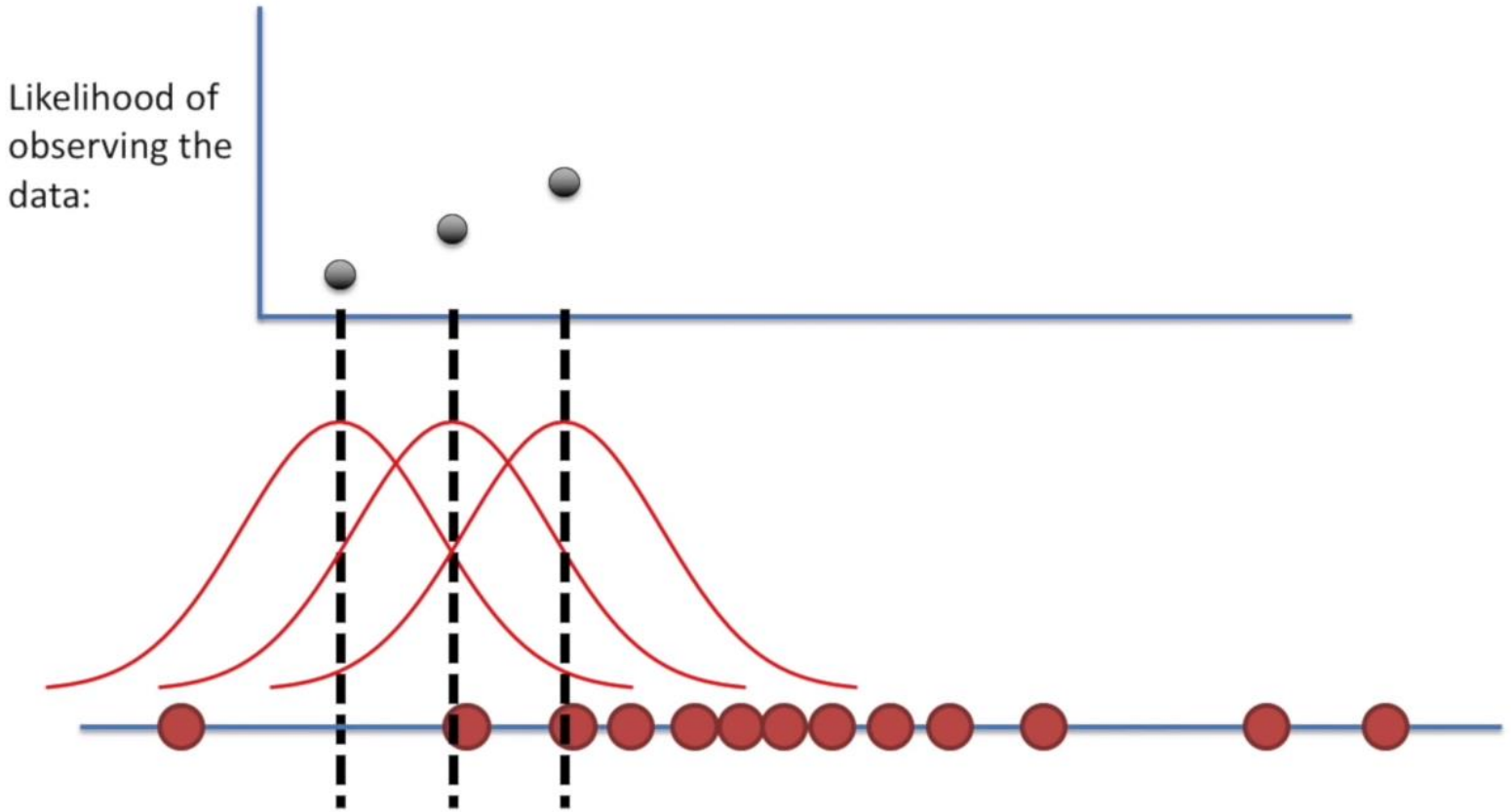
L(mean = 34 and standard deviation = 2.5 | mouse weighs 34 grams)

최대우도추정(MLE)

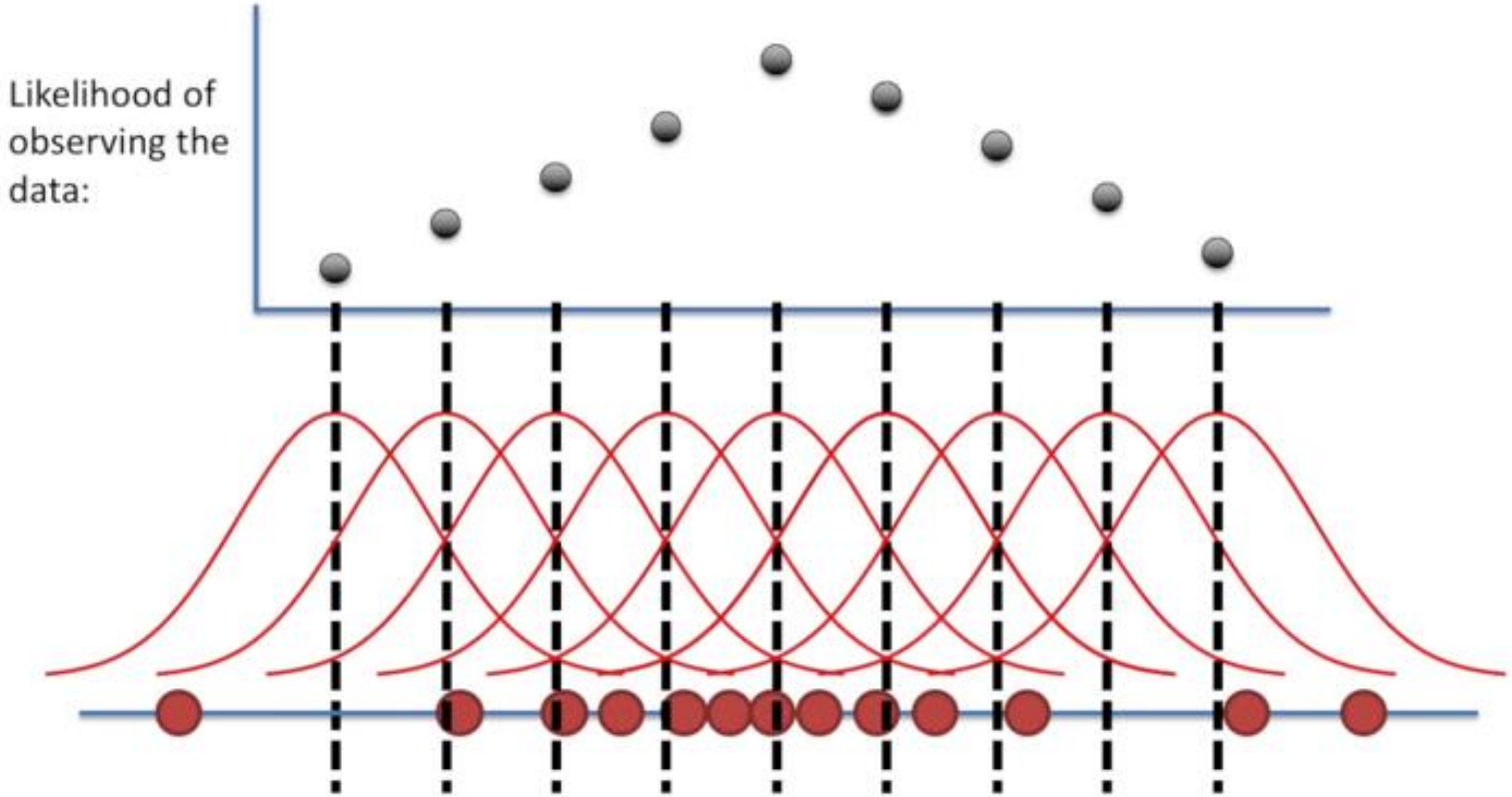
최대우도추정(MLE)


Likelihood of observing the data:
Location of the center of the distribution.

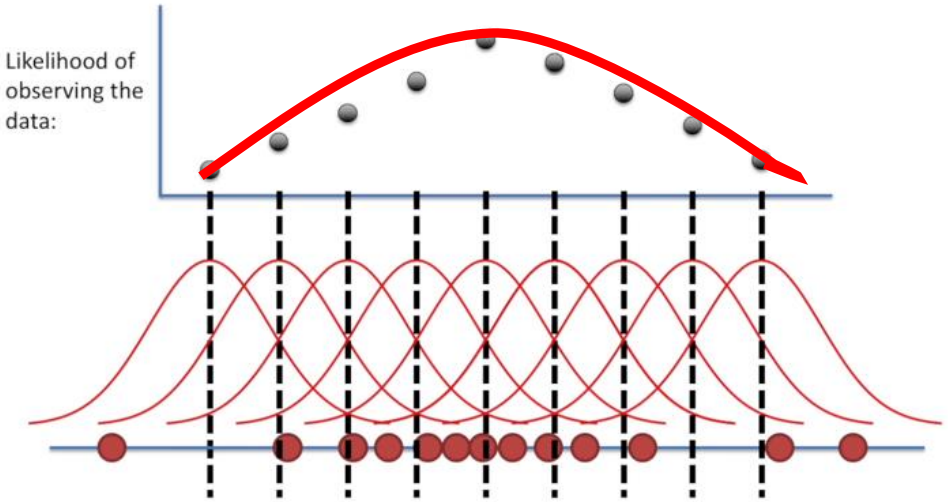
최대우도추정(MLE)

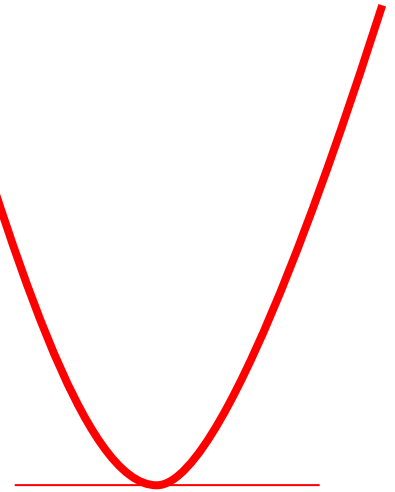


최대우도추정(MLE)



최대우도추정(MLE)

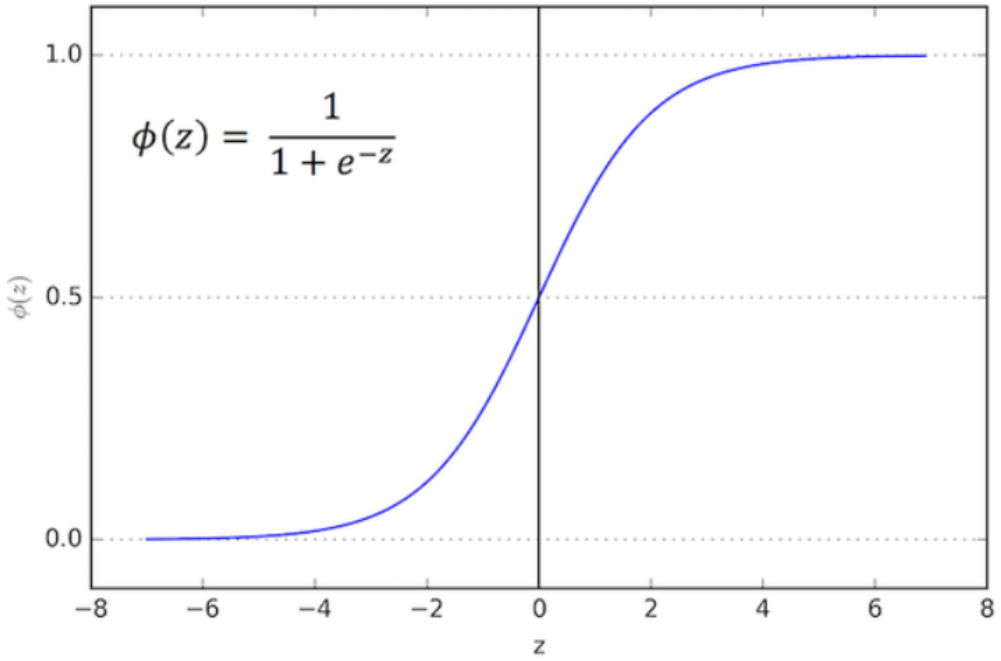




$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Unit 04 | 최대우도추정

최종 로지스틱 모델



최적 파라미터를 적합시킨 모델

$$\pi(X) = f(X) = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_k} x_k)}}$$

$$= \frac{1}{1 + e^{-\widehat{\beta} X}}$$

Unit 04 | 최대우도추정

Model Evaluation

Confusion Matrix

Predicted \ Actual		실제 정답 P	실제 정답 N
분류 결과	P	True Positive	False Positive
	N	False Negative	True Negative

1. True / False: 예측이 정확한가(T) 아닌가(F)?
2. Positive / Negative :
   1로 예측하면 Positive, 0으로 예측하면 Negative

Accuracy : 정확도

- 예측 결과가 실제와 얼마나 동일한지 측정
- 실제 분포가 편향(skewed) 되어 있는 경우 적합하지 않음

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

ex. Y = 질병 유무
질병이 없는 경우(Y=0)가 질병이 있는 경우(Y=1)보다 훨씬 많을 것!
이 때 분류 모형을 학습시키게 되면 Y=0일 때를 더 많이 학습하게 됨
-> 실제 데이터와 무관하게 Y=0이라고 예측할 확률이 커짐

즉, Accuracy는 TN, TP를 한번에 고려하므로,
TN은 높지만 TP가 낮은 경우는 고려하지 못하게 됨 !

Unit 04 | 최대우도추정

Precision and Recall

Actual

Confusion Matrix		실제 정답 P	실제 정답 N
Predicted 분류 결과	P	True Positive	False Positive
	N	False Negative	True Negative

Precision : 정밀도

- True라고 분류한 것 중에서 실제 True인 것의 비율

$$\text{Precision} = \frac{TP}{TP+FP}$$

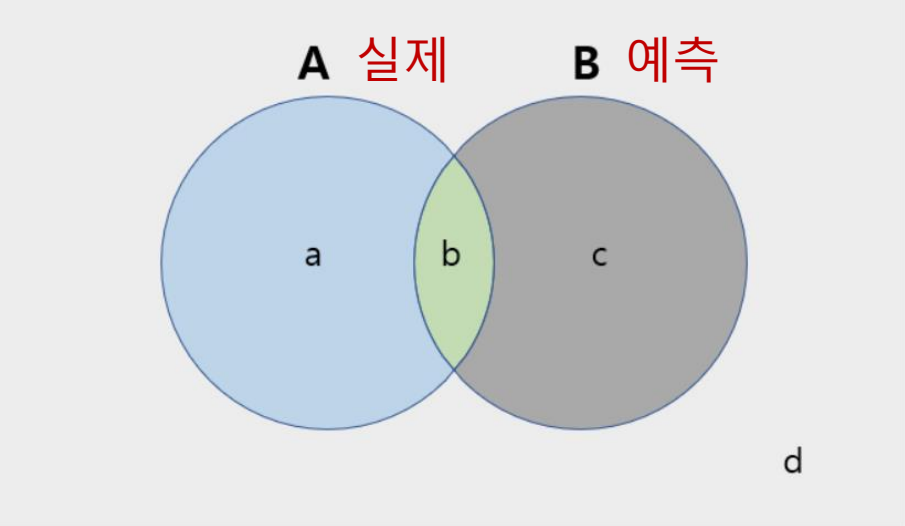
Recall (= sensitivity) : 재현율

- 실제 True인 것 중에서 True라고 예측한 것의 비율

$$\text{Recall} = \frac{TP}{TP+FN}$$

(cf) Precision과 Recall은 Trade-off 관계

-> 두 개의 값을 동시에 높일 수 없다!



$$\text{Precision} = \frac{b}{b+c} \quad , \quad \text{Recall} = \frac{b}{a+b}$$

a 부분이 c로 다 흡수된다면..?

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

$$\text{Precision} = \frac{20}{60} = 33.3\%$$

$$\text{Recall} = \frac{20}{50} = 40\%$$

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

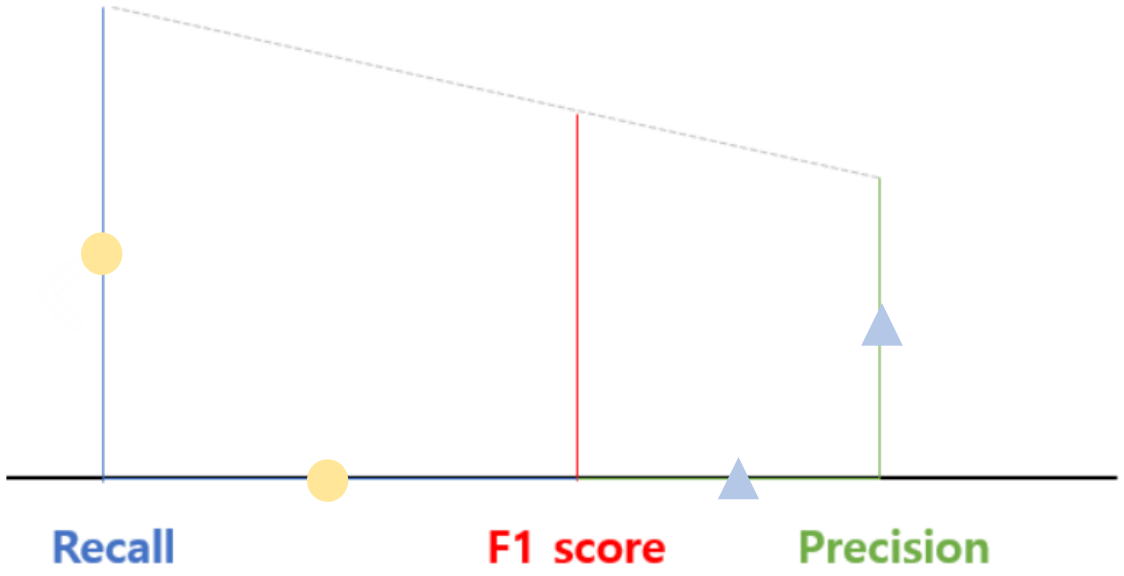
$$\text{Precision} = \frac{20}{100} = 20\%$$

$$\text{Recall} = \frac{20}{20} = 100\%$$

Unit 04 | 최대우도추정

F1 score

Precision과 Recall의 조화평균



F1 score

$$= 2 \times \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}}$$

$$= 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Unit 04 | 로지스틱 회귀분석

< 마무리 > 로지스틱 회귀분석

1. 범주형 변수 Y 분류
2. f(x) = 1 / (1+exp(-X*beta)) : 확률값 예측
3. Logit = log(Odds) = log(p/(1-p))
4. Beta1 = log(Odds)의 변화량
5. 최적의 가설 (beta) 구하기, MLE 이용
6. Recall, Precision, F1- Score 기준 Classification 성능개선

Assignment

<과제1> 행렬 구현
- LSE normal equation, MSE 구현 (Assignment1 파일에서 함수 구현하기)
- MLE 서술형 문제

<과제2> 회귀분석 : Used Car Price Prediction
- 배운 내용을 토대로 자유롭게 회귀분석과 회귀진단을 해주세요
- **해석을 상세하게 달아주세요 !**

<과제3> 로지스틱 회귀분석 : Credit Card Fraud Detection
1. sklearn 패키지를 사용해 로지스틱 회귀모형으로 데이터를 분석해 주세요
2. 성능지표를 계산하고 이에 대해 해석해 주세요
   - sklearn : mean accuracy, f1 score 등 다양한 성능지표 계산
   - confusion matrix : tp, fp, fn, tn 값을 통해 성능지표 계산
3. 어떤 성능지표를 기준으로 성능을 개선을 시도했고, 선택의 이유를 적어주세요.
- **해석을 상세하게 달아주세요 !**

Reference

<회귀분석>
투빅스 12기 이홍정님 강의자료, 투빅스 11기 심은선님 강의자료 / 투빅스 2기 김상진님 강의자료
이화여자대학교 통계학과 임용빈 교수님 강의
건국대학교 응용통계학과 유규상 교수님 강의 <회귀분석>
건국대학교 전자전기공학부 김원준 교수님 강의
Regression Analysis by Example edition 5 , Samprit Chatterjee/ Ali S. Hadi
<로지스틱 회귀분석>
투빅스 12기 이유진님 강의자료, 투빅스 11기 이영전님 강의자료
건국대학교 응용통계학과 유규상 교수님 강의 <경제자료분석>
건국대학교 전자전기공학부 김원준 교수님 강의
로지스틱 회귀분석에서 통계량의 이해 : https://nittaku.tistory.com/478
<최대우도 추정>
https://rk1993.tistory.com/entry/%EC%B5%9C%EB%8C%80%EC%9A%B0%EB%8F%84%EC%B6%94%EC%A0%95%EB%B2%95
https://jjangjjong.tistory.com/41

들어주셔서 감사합니다.

Unit 01 | 선형 회귀분석

최소제곱법 with 행렬

n x 1 n x (p+1) (p+1) x 1 n x 1

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

절편 변수1 변수p

road_name	road_bunji1	road_bunji2	Close_date	Close_result	point.y	point.x
해운대해변로	30.0	NaN	2018-06-14 00:00:00	배당	35.152717	129.137048
마린시티2로	33.0	NaN	2017-03-30 00:00:00	배당	35.158633	129.145068

$$Y = X\beta + \varepsilon$$

n : data 개수
p : feature 개수 (column / variable)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 & \beta_1 x_{11} & \cdots \beta_p x_{1p} \\ \beta_0 & \beta_1 x_{21} & \cdots \beta_p x_{2p} \\ \vdots & \vdots & \vdots \\ \beta_0 & \beta_1 x_{n1} & \cdots \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Unit 01 | 선형 회귀분석

최소제곱법 with 행렬

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i + \cdots + \beta_p x_i))^2$$

목적함수

$$= (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta)$$

↓ 최소화하므로 미분 = 0

$$\frac{\partial L}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

< Normal Equation > 정규방정식

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

Cross Entropy

- 분류에서의 학습을 위한 손실함수 = 입력값과 출력분포의 차이를 최소화

- Likelihood : $L(\mathbf{y},\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
- log-likelihood : $\log\big(L(p)\big) = \sum_{i=1}^{n}[y_i \log(p) + (1-y_i)\log(1-p)]$

- Cross-Entropy Loss : $-\log(L) = -\sum_{i=1}^{n}[y_i \log(p) + (1-y_i)\log(1-p)]$

- Cross-Entropy Loss Minimize = log-likelihood Function Maximize

Unit 03 | 로지스틱 회귀분석

Multiclass 범주에서의 로지스틱 회귀

https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/

- Binary Classification : log(Odds)
- Multiclass Classification : Baseline logit model

$$\log \frac{P(Y=1|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_1^T \vec{x}$$

$$\log \frac{P(Y=2|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_2^T \vec{x}$$

Y=3을 기준으로 하는 baseline logit model
P(Y=3) = 1 – P(Y=1) – P(Y=2)
-> 두 개의 계수 추정만 이루어지게 됨 !

앞에서 했던 것처럼 로그 확률비를 확률의 형태로 변환하고, 일반화된 형태를 취하면, 다음과 같은 형태를 보임

$$P(Y=c) = \frac{e^{\vec{\beta_c}^T \vec{x}}}{\sum_{k=1}^{K} e^{\vec{\beta_k}^T \vec{x}}}$$

c번째 범주에 속할 확률
⇔
Neural Network의 활성화 함수로 쓰이는
Softmax 함수와 동일한 형태 !

MLE in 로지스틱 회귀

- 관측값 y_i에서의 확률 분포 : $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1, 2, ..., n$

- Likelihood Function : $L(\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- log-likelihood : $\ln L(\mathbf{y}, \boldsymbol{\beta}) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i)$ →

MLE in 로지스틱 회귀

$$\ln L = \sum y_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) - \sum \ln(1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k})$$

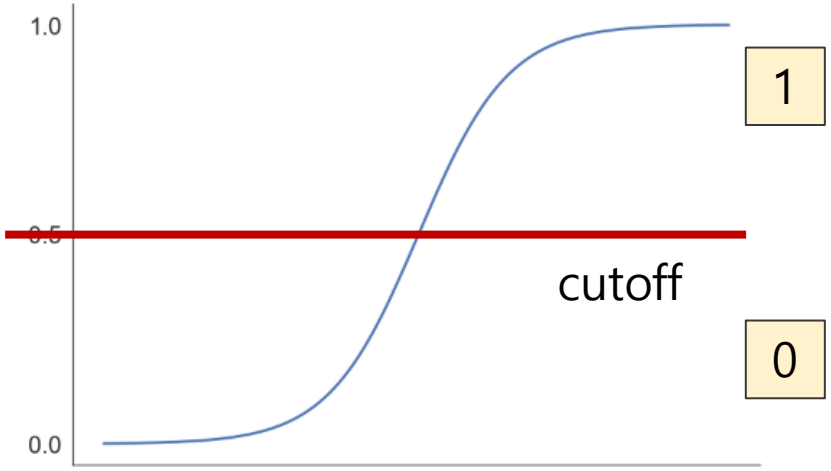
(cf) $\log(Odds) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ & $\pi_i = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$

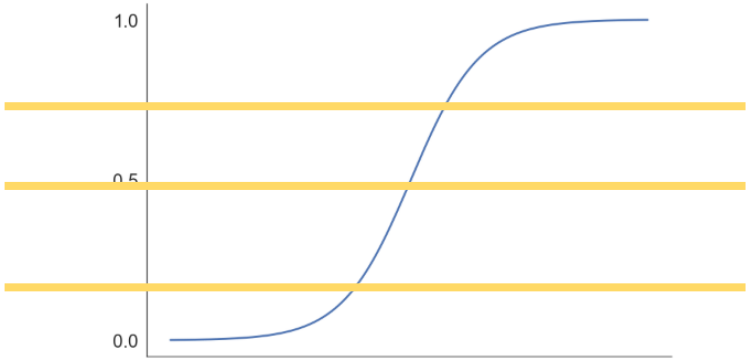
→ log-likelihood function이 최대가 되는 파라미터 β 결정 !

- log-likelihood 함수는 비선형 함수이므로, 선형회귀 모델처럼 명시적인 해가 존재하지 않음
- 따라서 Gradient Descent 등의 수치 최적화 알고리즘을 이용해 해를 구합니다 !

Cutoff (=Threshold)

- Classification 을 위한 기준값
- 로지스틱 함수로부터 구한 성공확률이 cutoff 이상이면 1 / cutoff 이하이면 0 으로 분류

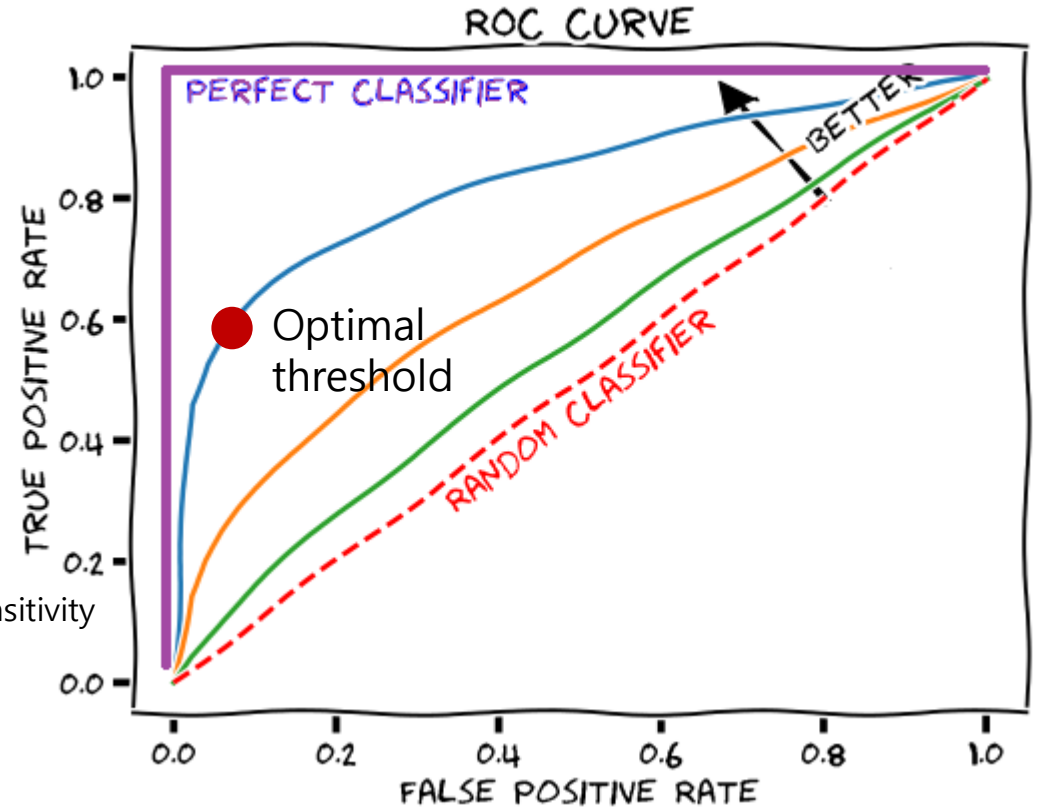




✓ 사전 확률을 고려한 cutoff
✓ 검증 데이터의 성능을 최대화하는 cutoff

Unit 04 | 로지스틱 회귀분석

ROC Curve



여러 cutoff value 값을 기준으로
-> confusion matrix에서 sensitivity, specificity 계산
-> 값을 기준으로 그림을 그린 것

AUC (=Area Under Curve)
= ROC curve의 넓이 ($0.5 \leq AUC \leq 1$)
= 값이 클수록 모델의 성능이 좋다

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad , \quad \text{Specificity} = \frac{TN}{FP+TN}$$