



세션 로딩 중



회귀모델



BADA 세션

Contents

1. 개요
 - 머신러닝
 - 회귀분석의 목적과 활용
2. 단순 선형 회귀 모형
 - 공분산과 상관계수
 - 회귀 계수 추정
 - 최소제곱법
3. 다중 선형 회귀 모형
 - 다중공선성
 - 변수선택법
 - 모델 평가 및 해석
4. 그 외 회귀모형
 - Underfitting, Overfitting
 - Lasso 회귀 (L1 규제)
 - Ridge 회귀 (L2 규제)
 - Elastic net

Machine Learning

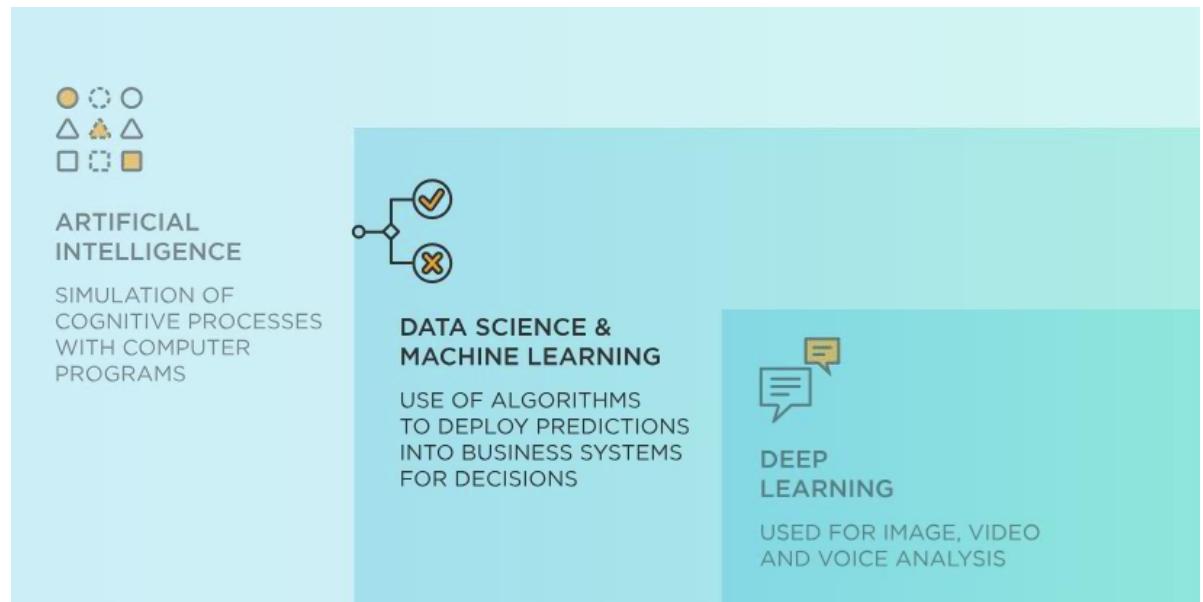
머신러닝

정의

- 특정한 과업(Task)을 달성하기 위한 경험(Experience)이 축적될수록 과업 수행의 성능(Performance)이 향상되는 컴퓨터 프로그램 또는 에이전트를 개발하는 것
- ✓ Data → Methodology (Model) → Results

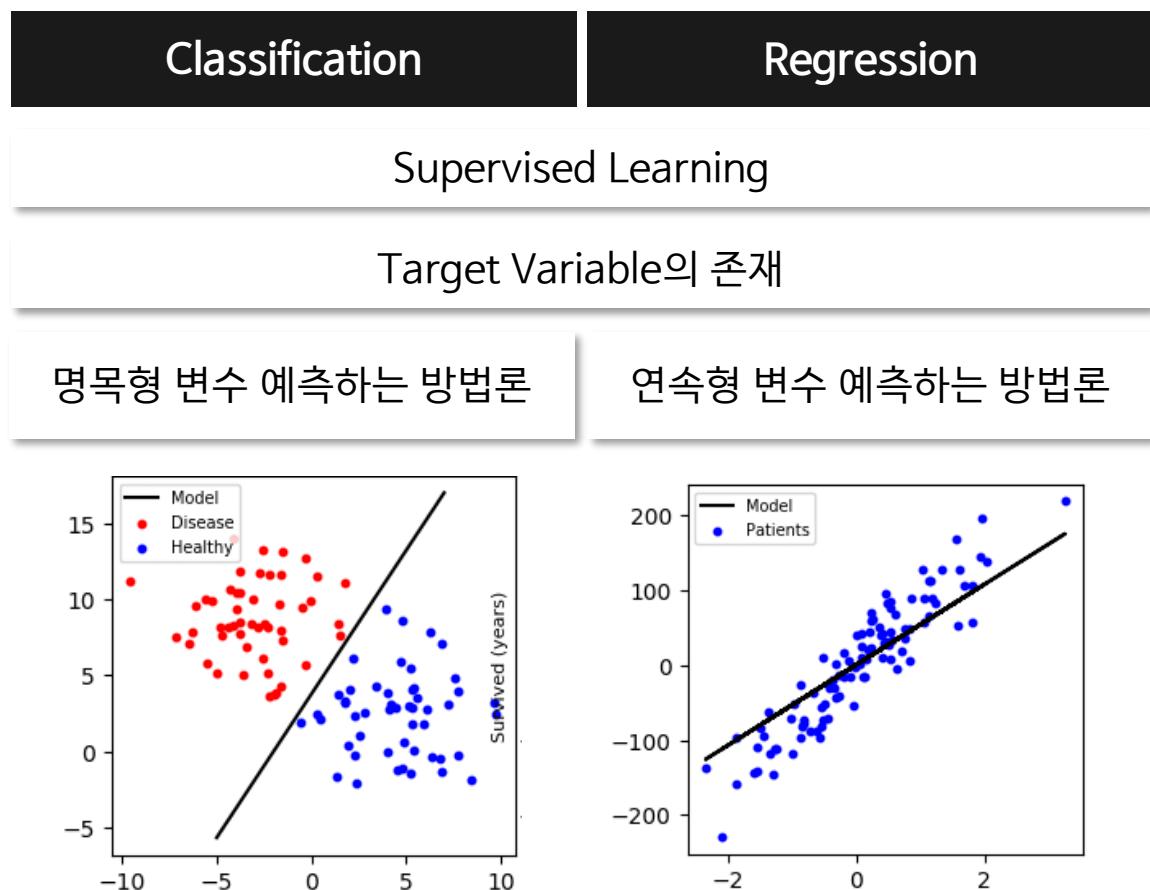
인공지능 ⊂ 머신러닝 ⊂ 딥러닝

- 인공지능이 가장 상위 개념이며,
딥러닝은 기계학습의 한 종류임



머신러닝의 종류

머신러닝은 1) Supervised Learning과 2) Unsupervised Learning으로 분류



› Classification	
공부시간	합격 여부 (합격, 불합격)
품종, 산도, 당도, 지역	와인의 등급
메일 발신인, 제목, 내용	스팸 메일 여부
› Regression	
온실 기체량	기온 변화
온도	음료 판매량
역세권, 조망	집 값

Regression

회귀분석

정의

- 독립변수와 종속변수 간의 함수관계를 규명하는 통계적 방법
- ✓ X 의 분포를 분석한 후, 종속변수 Y 의 값을 예측하는 것

독립변수

- 설명변수, 예측변수
- 어떤 효과를 관찰하기 위해 조작하는 변수

종속변수

- 반응변수, 결과변수
- 독립변수의 효과를 측정하는 대상

온도

음료 판매량 (quantitative)

- ✓ *Fit a linear relationship between a quantitative dependent variable Y and a set of predictors $X_1, X_2 \dots X_p$*

- 환경 분야

: 강 유역 주위의 토지 사용이 평균 질소 농도 (mg/liter) 단위의 수질 오염에 어떻게 영향을 주었는가?

- 의료서비스 분야

: 보건시설의 특성이 환자 치료 수익에 영향을 주는가?

- 컴퓨터 판매 • 수리 회사

: 컴퓨터 수리시간 (Y)과 수리 부품의 수 (X)

회귀분석 Case Study

삼성전자 Global CS센터 품질혁신팀_리뷰를 활용한 고객 품질 만족도 지수 개발: TV 사례연구

목적

- 기업간 치열한 경쟁으로 고객만족의 중요성 대두
- 설문조사의 한계 (사전설계된 질문지, 고객의 태도, 시간과 비용, 급변하는 고객 심리 반영 X)
- 전자상거래 활성화로 온라인에 고객리뷰 축적
→ 리뷰를 새로운 품질 만족도 지수에 활용하고자 함.

단계

1. 회귀분석을 활용해 별점과 리뷰 텍스트에서 특성별 중요도를 산출하고 제품의 어떤 특성이 고객만족에 주요한 영향을 주는지 파악한다.
2. 리뷰 텍스트를 특성별로 분류해 감성분석함으로써 특성별 긍정률을 산출한다.
3. 특성별 중요도와 긍정률을 가중합하여 품질 만족도 지수를 산출한다.

✓ Multiple Linear Regression

품질 특성의 긍정, 부정 언급

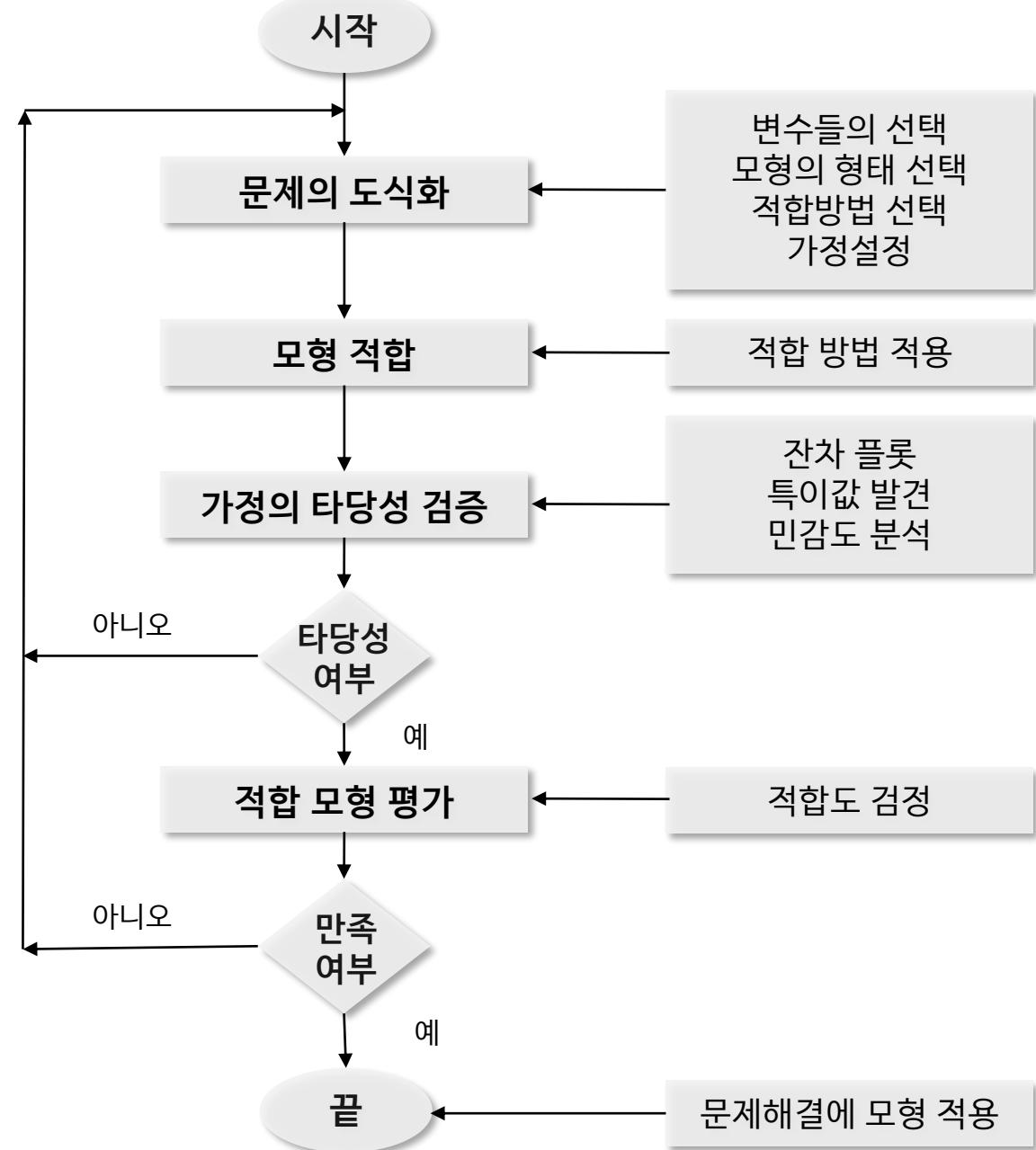
별점

$$\text{StarRate}_j = \beta_0 + \sum_{i=1}^k (\text{Coef}_i^{(p)} \text{Cnt}_{ij}^{(p)} + \text{Coef}_i^{(n)} \text{Cnt}_{ij}^{(n)}) + \sum_{i=k+1}^c (\text{Coef}_i^{(p)} \text{Cnt}_{ij}^{(p)} + \text{Coef}_i^{(n)} \text{Cnt}_{ij}^{(n)}) + \varepsilon_j$$

Attributes	2020	
	Negative	Positive
Picture quality	-0.035	0.119
Build quality	-0.062	0.014
Simplicity	-0.004	0.028

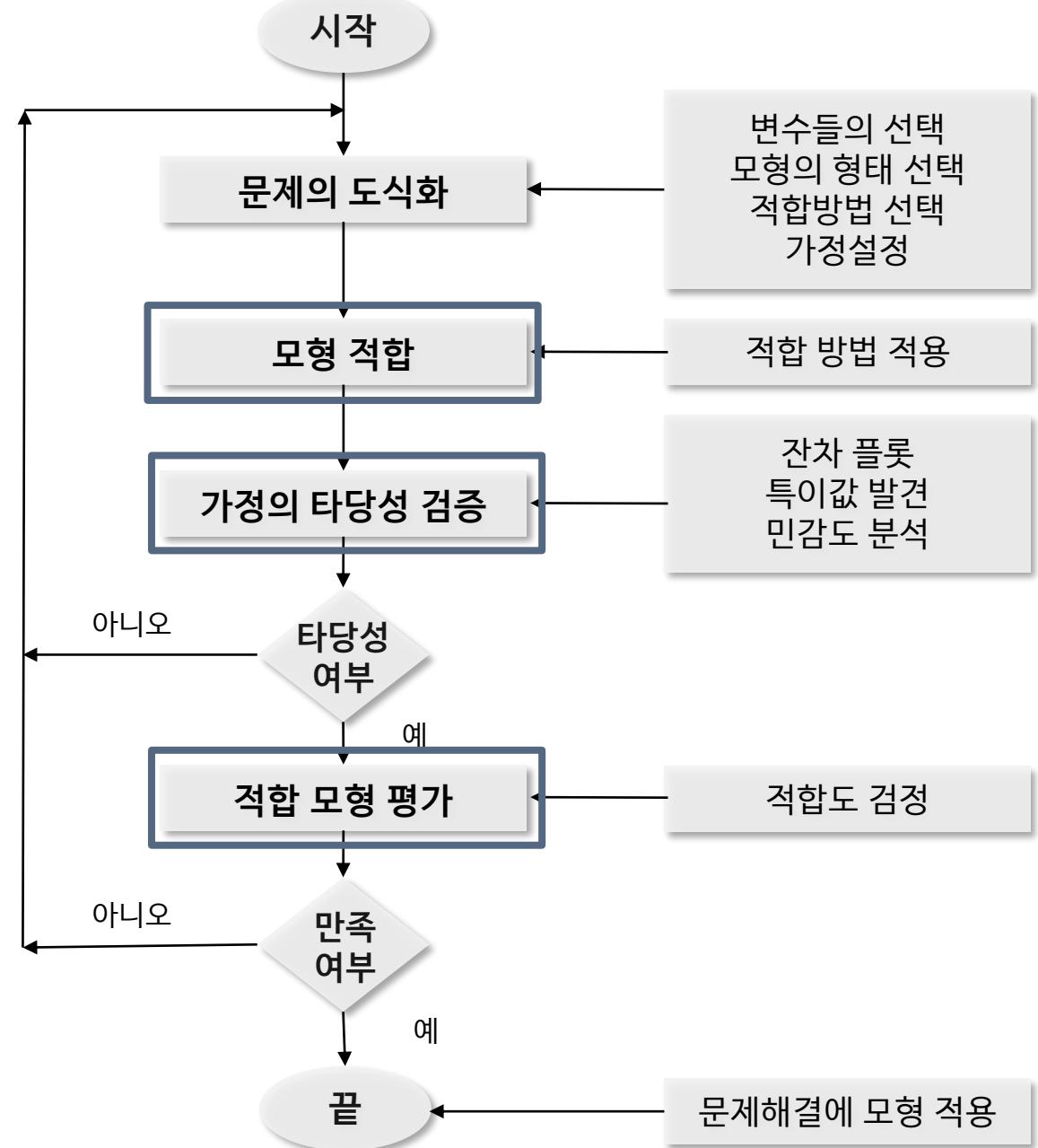
회귀분석 과정

반복적 회귀 과정의 플로우 차트



회귀분석 과정

반복적 회귀 과정의 플로우 차트



Simple Regression Model

분산, 공분산

분산

- $\sigma^2 = \text{Var}(x) = E[(X-\mu)^2]$
- 분산이 크면 데이터 변화가 큼
- 분산이 적으면 데이터 변화가 적음

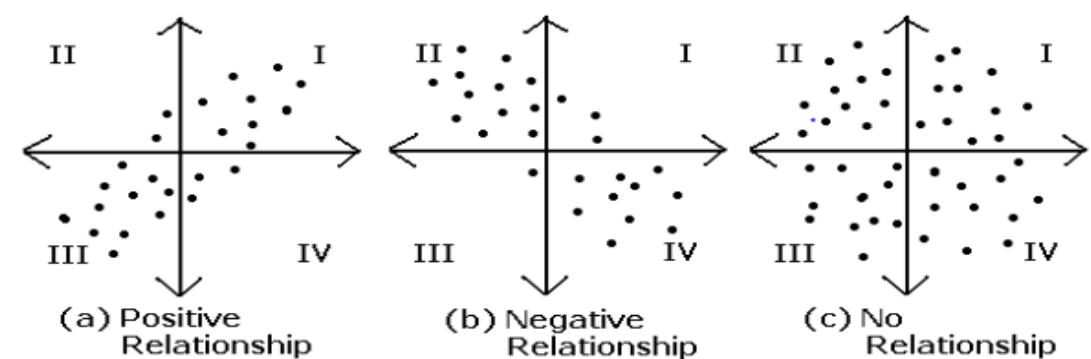
공분산

- $\text{Cov}(X, Y) = E[(X - \mu_1)(Y - \mu_2)] = E(XY) - \mu_1\mu_2$
 $= E(XY) - E(X)E(Y)$
(μ_1 : x의 평균 μ_2 = y의 평균)
- 2개의 변수들의 상관 정도/의존성/유사성을 나타내는 값

〉 공분산의 문제점

Ex. 키와 몸무게

- 공분산의 단위는 각 x와 y의 곱이다
- 각 x와 y들의 크기가 크면 공분산의 크기가 크게 나오는 문제 발생



출처 : <https://destrudo.tistory.com/15>

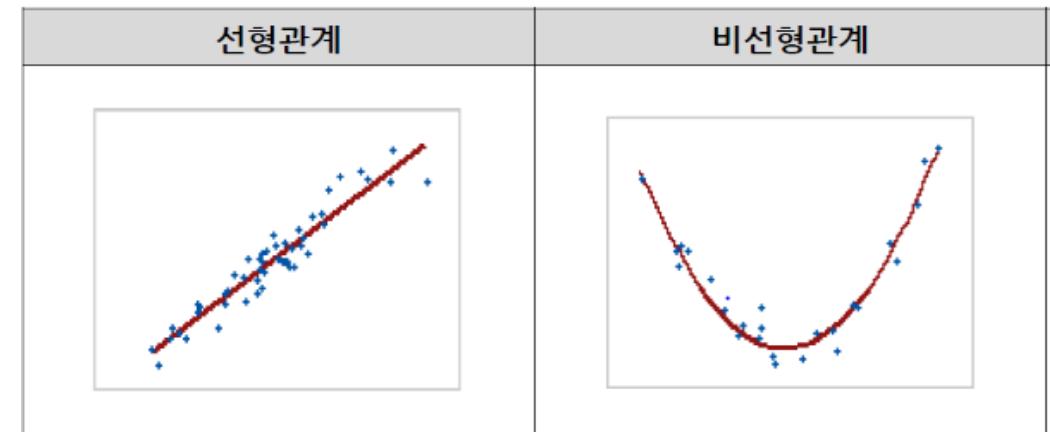
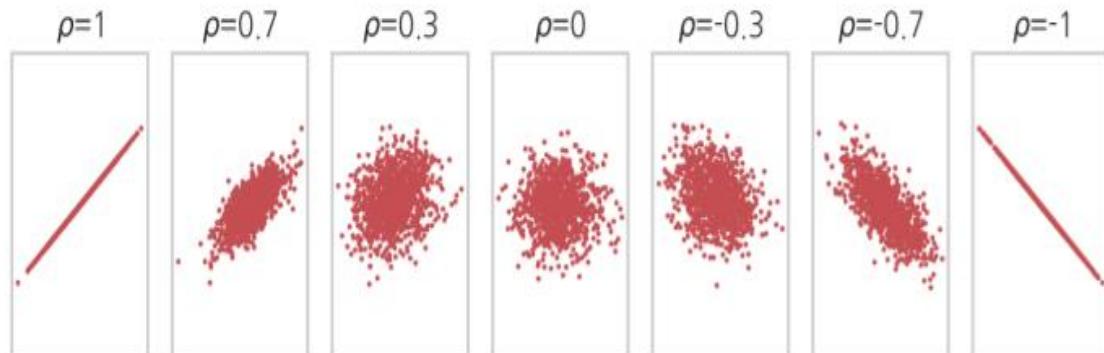
상관계수

공분산의 문제점을 해결하고자 상관계수를 사용

상관계수

- $\text{Corr}(X, Y) = \text{Cov}(X, Y) / \sigma(X) \sigma(Y)$
- 확률 변수의 절대적 크기에 영향을 받지 않도록 공분산을 단위화 시킨 것
- $-1 \leq \text{corr}(X, Y) \leq 1$

상관계수와 스캐터 플롯의 모양



› 상관계수의 한계

- 비선형 관계의 상관계수는 거의 0에 가까움
- 상관관계는 ‘선형적’ 관계만 나타내기 때문에 상관관계가 0이라는 것은 선형적 관계만 없다는 것일 뿐 관계가 없다고 할 수 없음

상관에서 회귀로

〉 공분산과 상관계수 역할

- 탐색적 데이터 분석 (EDA)에서 매우 중요!
- WHY? 통계적 도구로 사용하여 변수들 간의 선형적 관계를 알아보고 이에 따른 모델의 방향성을 설정하고 변수 선택 등 다양한 역할

상관에서 회귀로의 확장

상관계수

: 두 변수 간 선형적 관계의 강도와 방향을 나타내는 통계적 도구
But, 상관계수만으로는 이 관계가 통계적 유의성, 예측불가

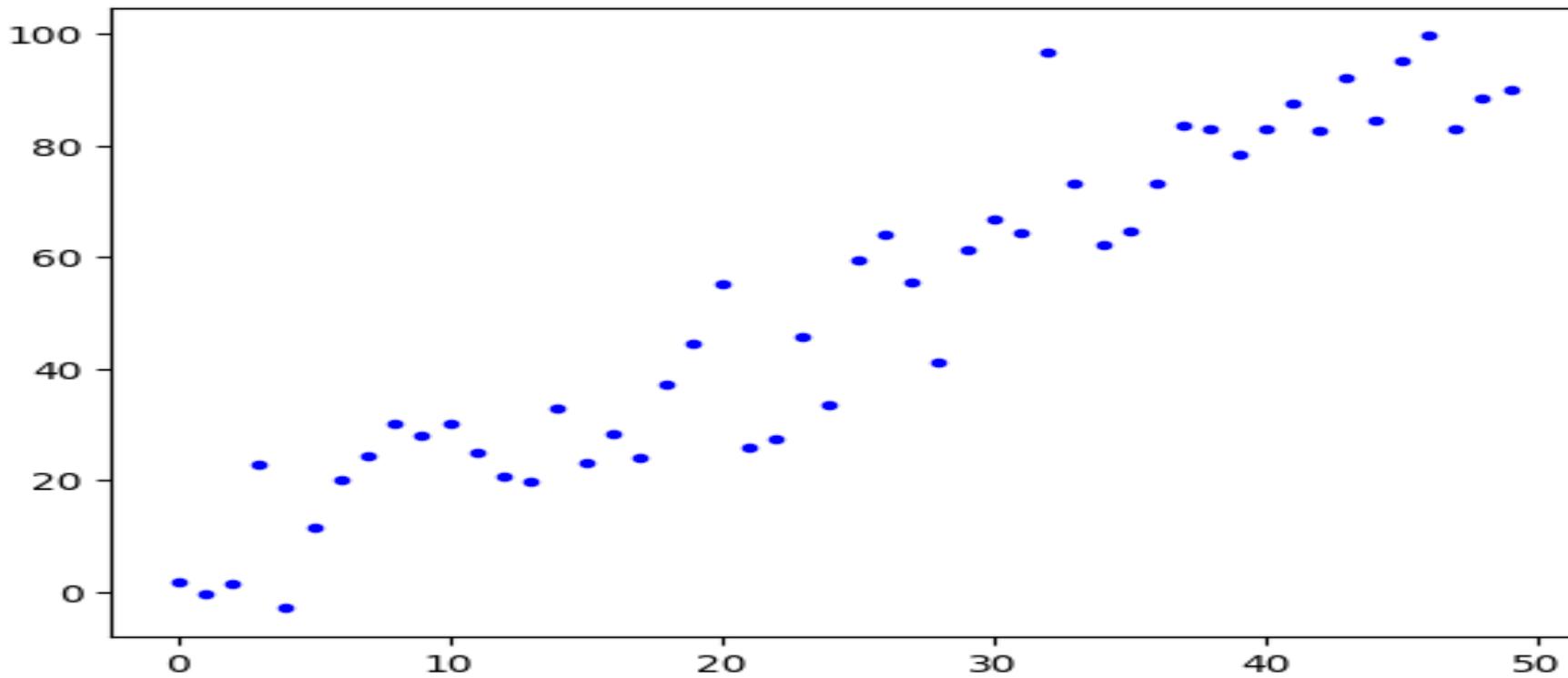
회귀

: 두 변수 간의 관계를 보다 구체적으로 규명

- 독립변수들과 종속 변수에 미치는 영향을 수치적으로 표현하고, 이러한 관계를 기반으로 예측 가능

✓ 주의! 여기서의 관계란 상관분석에서 본 선형적인 연관성을 뜻함

데이터를 어떻게 표현하는 것이 좋을까?



단순 선형회귀

단순 선형 회귀식

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

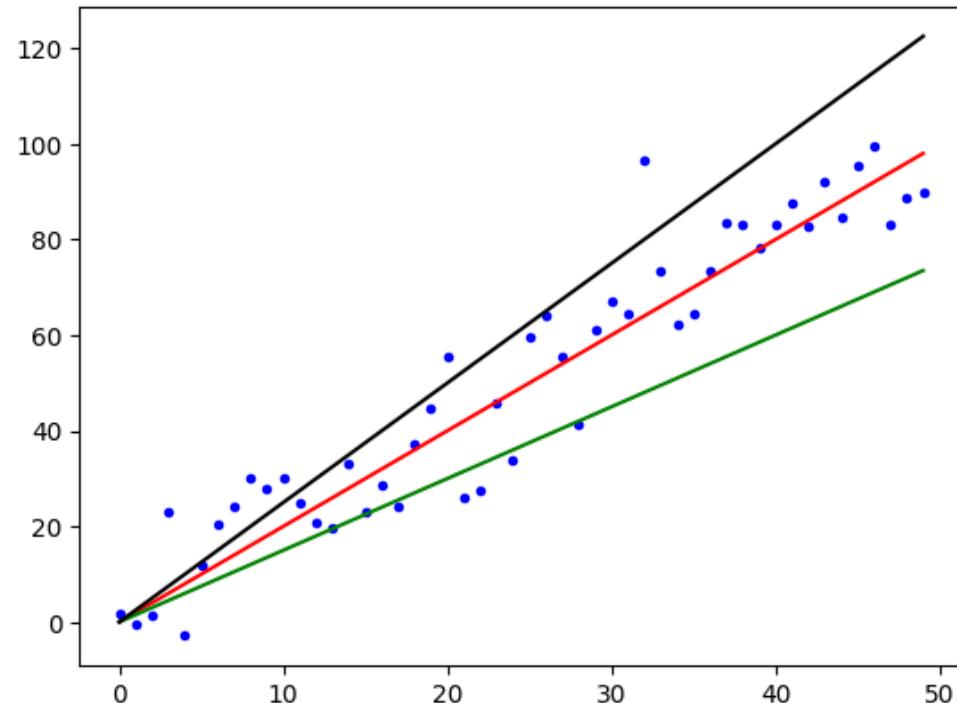
- Y: 종속 변수 (예측하려는 값)
- X: 독립 변수 (Y를 예측하는 데 사용되는 변수)
- β_0 : Y 축의 회귀 절편 (절편 또는 상수항)
- β_1 : 회귀 계수 (독립 변수 X의 기울기)
- ε : 오차 항 (모델로 설명되지 않는 잔차 또는 오차)

✓ 선형회귀에서는 회귀계수와 회귀절편에 주목해야한다!

- β_0 (Y 축의 회귀 절편)
: X가 0일 때 ($X = 0$)의 Y 값,
즉 X와 Y의 관계가 어떤 위치에서 시작하는지를 나타냄.
- β_1 (회귀 계수)
: X와 Y 간의 선형 관계의 기울기로
X가 1 증가할 때 Y가 얼마나 증가하는지
또는 감소하는지를 나타냄.

단순 선형회귀

검은색, 초록색, 빨간색 중 어떤 선이 더 잘 표현하였는가?



회귀분석의 목적

- 실제 데이터와 예측한 데이터 값을 최소화하는 회귀식을 찾는 것.

최소제곱법

- 데이터와 회귀선 사이의 거리 (에러)의 제곱을 최소화

최대우도법

- 데이터를 가장 잘 설명하는 확률분포 모델의 파라미터 찾는 법

최소제곱법

정의

- 근사적으로 구하려는 해와 실제 해의 오차의 제곱의 합(SS)이 최소가 되는 해를 구하는 방법
- 즉, $\varepsilon^2 = (\hat{y} - y)^2$ 를 최소로 만드는 해 찾기!

목표: $\text{Min } \sum_{i=1}^n (y - \hat{y})^2$ $\hat{y} = \beta_0 + \beta_1 x_i$

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

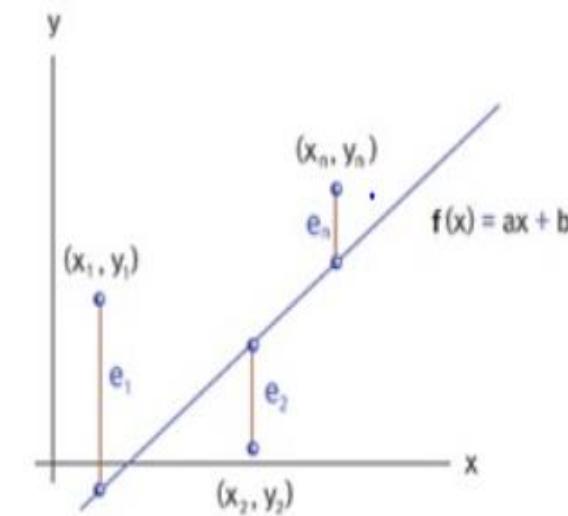
$$\frac{d}{d\beta_0} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (2)$$

$$\beta_0 = \sum_{i=1}^n y_i / n - \beta_1 * \sum_{i=1}^n x_i / n = \bar{Y} - \beta_1 \bar{X} \quad (3)$$

$$\beta_1 = \sum_{i=1}^n x_i (y_i - \bar{Y}) / \sum_{i=1}^n x_i (x_i - \bar{X})$$

$$\beta_1 = \text{cov}(x,y) / \text{var}(x)$$

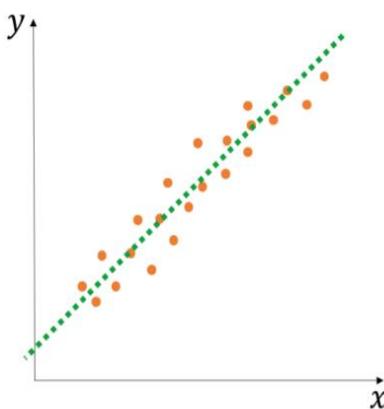
$$\beta_0 = \bar{Y} - \bar{X} \beta_1$$



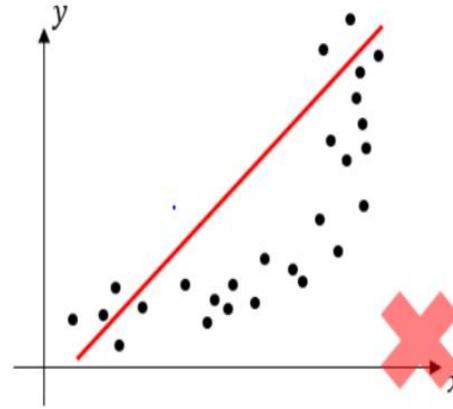
✓ 오차의 제곱합 최소화!

회귀분석의 가정

1. 선형성



> 선형성 만족 O



> 선형성 만족 X

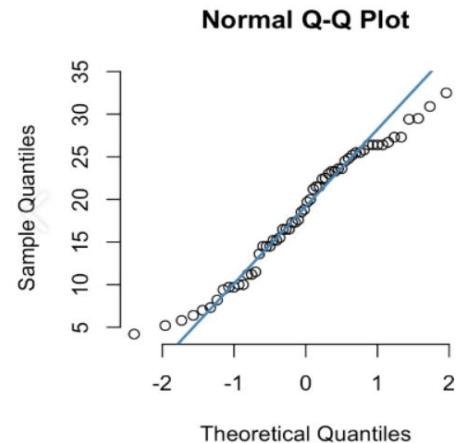
가정

: ‘종속변수와 독립변수 간에 선형관계가 존재한다’

가정을 만족하지 못하는 경우

- 데이터가 복잡한 비선형 패턴을 보이는 경우
- 데이터가 지수적으로 증가하거나 감소하는 경우

2. 잔차 정규성



- ✓ Q-Q plot 해석
점들이 대략적으로 직선에
가까이 배열되어 있으면 잔차가
정규 분포를 따른다고 볼 수 있음

가정

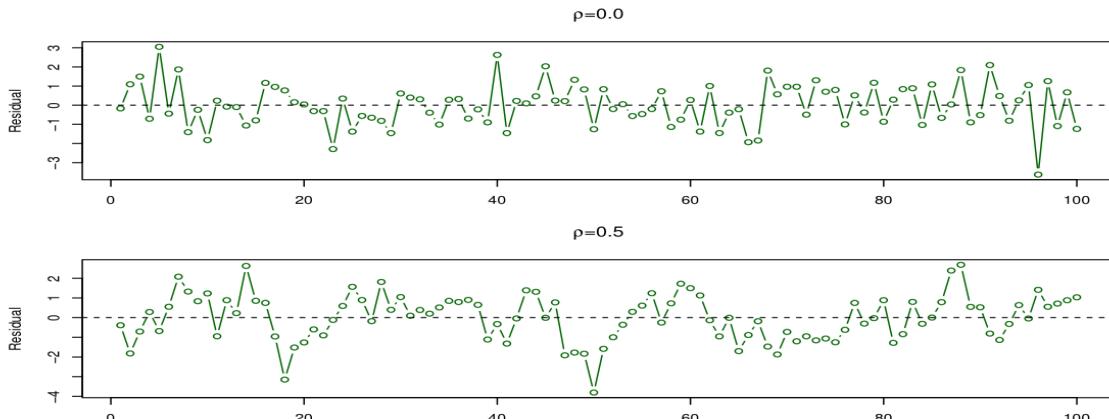
: ‘모델의 잔차가 정규분포를 따른다’

가정을 만족하지 못하는 경우

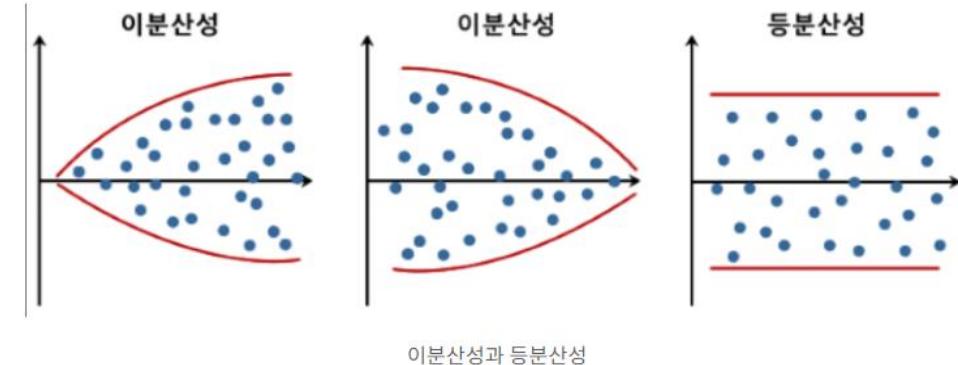
모델의 신뢰구간과 가설검정 부정확

회귀분석의 가정

3. 잔차 독립성 (시계열 데이터의 경우 유의)



4. 잔차 등분산성



가정

: ‘회귀 모델의 잔차가 서로 독립적이어야 한다.’

가정을 만족하지 못하는 경우

- 중요변수 누락 (중요변수의 패턴이 잔차에 나타날 가능성)
- 부적절한 모델구조 (비선형인데 선형하는 경우)
- 시계열 데이터 (계절성 및 추세)

가정

: ‘회귀 모델의 잔차가 모든 수준의 독립 변수에 대해 일정한 분산을 가져야 한다’

가정을 만족하지 못하는 경우

- 모델의 부적합, 극단값 또는 이상치 존재
- 중요 변수 누락, 비선형 관계

회귀분석 결과해석

OLS Regression Results						
Dep. Variable:	Y	R-squared:	0.881			
Model:	OLS	Adj. R-squared:	0.879			
Method:	Least Squares	F-statistic:	356.6			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24			
Time:	13:35:41	Log-Likelihood:	-188.98			
No. Observations:	50	AIC:	382.0			
Df Residuals:	48	BIC:	385.8			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.8571	3.014	0.284	0.777	-5.203	6.917
X	2.0018	0.106	18.885	0.000	1.789	2.215
Omnibus:	0.112	Durbin-Watson:	1.974			
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173			
Skew:	-0.102	Prob(JB):	0.917			
Kurtosis:	2.798	Cond. No.	56.1			

No. Observations

- 회귀분석에 사용된 데이터의 총 수

Df Residuals

- Df Residuals**=관측치 수 – (독립 변수의 수+1)
- 잔차의 자유도
- 높은 자유도는 모델이 데이터에 더 유연하게 적합될 수 있음을 의미

Df Model

- Df Model**= 모델에 의해 추정되는 독립 변수의 수
- 모델의 자유도
- 모델의 자유도는 모델의 복잡성을 나타내고 많은 독립 변수를 포함하는 모델은 더 높은 자유도를 가지며 모델이 데이터에 더 유연하게 적합될 수 있음을 의미
But 과적합의 위험도 증가

회귀분석 결과해석

OLS Regression Results						
Dep. Variable:	Y	R-squared:	0.881			
Model:	OLS	Adj. R-squared:	0.879			
Method:	Least Squares	F-statistic:	356.6			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24			
Time:	13:35:41	Log-Likelihood:	-188.98			
No. Observations:	50	AIC:	382.0			
Df Residuals:	48	BIC:	385.8			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.8571	3.014	0.284	0.777	-5.203	6.917
X	2.0018	0.106	18.885	0.000	1.789	2.215
Omnibus:	0.112	Durbin-Watson:	1.974			
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173			
Skew:	-0.102	Prob(JB):	0.917			
Kurtosis:	2.798	Cond. No.	56.1			

Model

- 분석의 특정 유형

Ex. 단순 선형 회귀, 다중 선형 회귀, 로지스틱 회귀
다항회귀 등 어떤 분석 유형인가?

Method

- 모델의 파라미터 추정하는 방법

Ex. 최소제곱법(LS), 최대우도법(MLE), 릿지(L2),
라쏘(L1)

회귀분석 결과해석

OLS Regression Results						
Dep. Variable:	Y	R-squared:	0.881			
Model:	OLS	Adj. R-squared:	0.879			
Method:	Least Squares	F-statistic:	356.6			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24			
Time:	13:35:41	Log-Likelihood:	-188.98			
No. Observations:	50	AIC:	382.0			
Df Residuals:	48	BIC:	385.8			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.8571	3.014	0.284	0.777	-5.203	6.917
X	2.0018	0.106	18.885	0.000	1.789	2.215
Omnibus:	0.112	Durbin-Watson:	1.974			
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173			
Skew:	-0.102	Prob(JB):	0.917			
Kurtosis:	2.798	Cond. No.	56.1			

결정계수

- 독립변수와 종속변수 간의 함수관계를 규명
 - X 의 분포를 분석한 후, 종속변수 Y 의 값을 예측
- ✓ 모델의 설명력을 나타내는 값으로, 종속 변수의 분산을 독립 변수가 얼마나 잘 높을수록 모델이 데이터를 잘 설명

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$SS_{\text{res}} = \sum(y - \hat{y})^2 = \sum(\text{잔차})^2$$

$$SS_{\text{tot}} = \sum(y - \bar{y})^2 = \sum(\text{편차})^2$$

회귀분석 결과해석

OLS Regression Results						
Dep. Variable:	Y	R-squared:	0.881			
Model:	OLS	Adj. R-squared:	0.879			
Method:	Least Squares	F-statistic:	356.6			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24			
Time:	13:35:41	Log-Likelihood:	-188.98			
No. Observations:	50	AIC:	382.0			
Df Residuals:	48	BIC:	385.8			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
const	0.8571	3.014	0.284	0.777	-5.203	6.917
X	2.0018	0.106	18.885	0.000	1.789	2.215
=====						
Omnibus:	0.112	Durbin-Watson:	1.974			
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173			
Skew:	-0.102	Prob(JB):	0.917			
Kurtosis:	2.798	Cond. No.	56.1			
=====						

Dep. Variable

- 회귀분석의 종속변수 (Y)
- Dependent Variable

Adj. R-squared

- 독립 변수의 수와 데이터 점의 수를 고려하여 조정된 R제곱 값

Why? X 변수가 많아질수록 R-squared 값이 설명력과 관계없이 값이 커지는 단점 존재

회귀분석 결과해석

OLS Regression Results									
Dep. Variable:	Y	R-squared:	0.881						
Model:	OLS	Adj. R-squared:	0.879						
Method:	Least Squares	F-statistic:	356.6						
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24						
Time:	13:35:41	Log-Likelihood:	-188.98						
No. Observations:	50	AIC:	382.0						
Df Residuals:	48	BIC:	385.8						
Df Model:	1								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
const	0.8571	3.014	0.284	0.777	-5.203	6.917			
X	2.0018	0.106	18.885	0.000	1.789	2.215			
Omnibus:	0.112	Durbin-Watson:	1.974						
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173						
Skew:	-0.102	Prob(JB):	0.917						
Kurtosis:	2.798	Cond. No.	56.1						

Coef (계수)

- 회귀 모델의 계수, 절편 (Intercept)과 X의 계수

Intercept (절편): X가 0일 때의 예측값

X의 계수: X가 1 단위 증가할 때 Y의 변화

P>|t| (p-value)

- 독립 변수의 유의성을 나타냄
- 작을수록 해당 독립 변수가 종속 변수에 유의미한 영향을 미친다는 것 의미
- 보통 0.05보다 작은 p-값을 가지는 독립 변수는 유의미한 변수

회귀분석 결과해석

OLS Regression Results						
Dep. Variable:	Y	R-squared:	0.881			
Model:	OLS	Adj. R-squared:	0.879			
Method:	Least Squares	F-statistic:	356.6			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	7.34e-24			
Time:	13:35:41	Log-Likelihood:	-188.98			
No. Observations:	50	AIC:	382.0			
Df Residuals:	48	BIC:	385.8			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	0.8571	3.014	0.284	0.777	-5.203	6.917
X	2.0018	0.106	18.885	0.000	1.789	2.215
Omnibus:	0.112	Durbin-Watson:	1.974			
Prob(Omnibus):	0.946	Jarque-Bera (JB):	0.173			
Skew:	-0.102	Prob(JB):	0.917			
Kurtosis:	2.798	Cond. No.	56.1			

[0.025, 0.975]

- 회귀 계수의 신뢰구간

- 넓은 신뢰구간: 추정치의 불확실성이 크다는 것을, 좁은 신뢰구간은 높은 정확도를 의미
- 신뢰구간 0 포함: 해당 변수의 영향력이 통계적으로 유의미하지 않다 해석

95% 신뢰구간 의미

'표본 100개의 신뢰구간 중에 95개는 실제 모집단의 통계량을 포함한다'

- 신뢰구간 식

$$\bar{x} \pm z_{\alpha} \times \frac{s}{\sqrt{n}} \quad (\text{신뢰수준 } 95\% z_{\alpha} = 1.96)$$

\bar{x} : 표본 평균

z_{α} : 신뢰수준에 해당하는 Z-값

N : 표본수

s : 표본 표준 오차.

Multiple Regression Model – Part 1

다중선형회귀(Multiple Regression Model)

정의

설명변수(독립변수)의 개수가 여러 개인 회귀 분석 모델

Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N + \epsilon$$

필요성

다른 변수를 통제한 채로 (=고정한 채로),
특정 독립변수가 종속변수에 미치는
영향만 따로 파악 가능!

$$Y = 0.3 + 0.01x_1 + 0.8x_2$$

회귀계수(베타) 계산

$$e_i = y_i(\text{실제 값}) - \hat{y}_i(\text{추정한 값})$$

$$\text{잔차의 제곱합 } SSE = \sum_{i=1}^n e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$

- 위의 SSE 식을 각각의 변수에 대해 편미분하여 나온 방정식들의 값을 모두 0으로 두고, 연립방정식을 풀어 최적 회귀계수를 추정!
- 혹은, 행렬을 통해서도 회귀계수를 계산할 수 있음
(기초수학 및 머신러닝 세션 참고)

회귀계수(베타)의 해석

단순선형회귀때와 동일하게, x 가 한 단위 변화했을 때, y 의 기대 변화량을 의미.
“다른 변수들이 갖는 값이 고정되어 있을 때, 특정 변수가 Y 에 미치는 영향”

다중선형회귀 : 회귀계수 계산(Recall)

행렬로 표현한 선형결합

$$\begin{array}{c}
 X \quad \times \quad \hat{\beta} \quad = \quad X\hat{\beta} \quad = \quad \hat{y} \quad - \quad y \quad = \quad \text{error} \\
 \begin{array}{|c|c|c|c|c|c|} \hline
 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ \hline
 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \hline
 1 & x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \\ \hline
 1 & x_{41} & x_{42} & x_{43} & x_{44} & x_{45} \\ \hline
 1 & x_{51} & x_{52} & x_{53} & x_{54} & x_{55} \\ \hline
 1 & x_{61} & x_{62} & x_{63} & x_{64} & x_{65} \\ \hline
 1 & x_{71} & x_{72} & x_{73} & x_{74} & x_{75} \\ \hline
 1 & x_{81} & x_{82} & x_{83} & x_{84} & x_{85} \\ \hline
 \end{array} \quad \quad \quad
 \begin{array}{|c|} \hline
 \hat{\beta}_0 \\ \hline
 \hat{\beta}_1 \\ \hline
 \hat{\beta}_2 \\ \hline
 \hat{\beta}_3 \\ \hline
 \hat{\beta}_4 \\ \hline
 \hat{\beta}_5 \\ \hline
 \end{array} \quad \quad \quad
 \begin{array}{|c|} \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \hat{\beta}_3 x_{13} + \hat{\beta}_4 x_{14} + \hat{\beta}_5 x_{15} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22} + \hat{\beta}_3 x_{23} + \hat{\beta}_4 x_{24} + \hat{\beta}_5 x_{25} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{31} + \hat{\beta}_2 x_{32} + \hat{\beta}_3 x_{33} + \hat{\beta}_4 x_{34} + \hat{\beta}_5 x_{35} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{41} + \hat{\beta}_2 x_{42} + \hat{\beta}_3 x_{43} + \hat{\beta}_4 x_{44} + \hat{\beta}_5 x_{45} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{51} + \hat{\beta}_2 x_{52} + \hat{\beta}_3 x_{53} + \hat{\beta}_4 x_{54} + \hat{\beta}_5 x_{55} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{61} + \hat{\beta}_2 x_{62} + \hat{\beta}_3 x_{63} + \hat{\beta}_4 x_{64} + \hat{\beta}_5 x_{65} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{71} + \hat{\beta}_2 x_{72} + \hat{\beta}_3 x_{73} + \hat{\beta}_4 x_{74} + \hat{\beta}_5 x_{75} \\ \hline
 \hat{\beta}_0 + \hat{\beta}_1 x_{81} + \hat{\beta}_2 x_{82} + \hat{\beta}_3 x_{83} + \hat{\beta}_4 x_{84} + \hat{\beta}_5 x_{85} \\ \hline
 \end{array} \quad \quad \quad
 \begin{array}{|c|} \hline
 \hat{y}_1 \\ \hline
 \hat{y}_2 \\ \hline
 \hat{y}_3 \\ \hline
 \hat{y}_4 \\ \hline
 \hat{y}_5 \\ \hline
 \hat{y}_6 \\ \hline
 \hat{y}_7 \\ \hline
 \hat{y}_8 \\ \hline
 \end{array} \quad \quad \quad
 \begin{array}{|c|} \hline
 y_1 \\ \hline
 y_2 \\ \hline
 y_3 \\ \hline
 y_4 \\ \hline
 y_5 \\ \hline
 y_6 \\ \hline
 y_7 \\ \hline
 y_8 \\ \hline
 \end{array} \quad \quad \quad
 \begin{array}{|c|} \hline
 \varepsilon_1 \\ \hline
 \varepsilon_2 \\ \hline
 \varepsilon_3 \\ \hline
 \varepsilon_4 \\ \hline
 \varepsilon_5 \\ \hline
 \varepsilon_6 \\ \hline
 \varepsilon_7 \\ \hline
 \varepsilon_8 \\ \hline
 \end{array}
 \end{array}$$

다중선형회귀 : 회귀계수 계산(Recall)

목적함수 정의(MSE)

$$\frac{1}{n} \sum_{i=1}^d (\text{실제값} - \text{예측값})^2 = \frac{1}{n} \sum_{i=1}^d (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^d \varepsilon_i^2$$

MSE를 최소화하는 $\hat{\beta}$ 를 찾는 것! 즉, $\min_{\hat{\beta}} (y - X\hat{\beta})^2$ 를 찾아야 함!

$$MSE = (y - X\hat{\beta})^2$$

회귀계수 계산

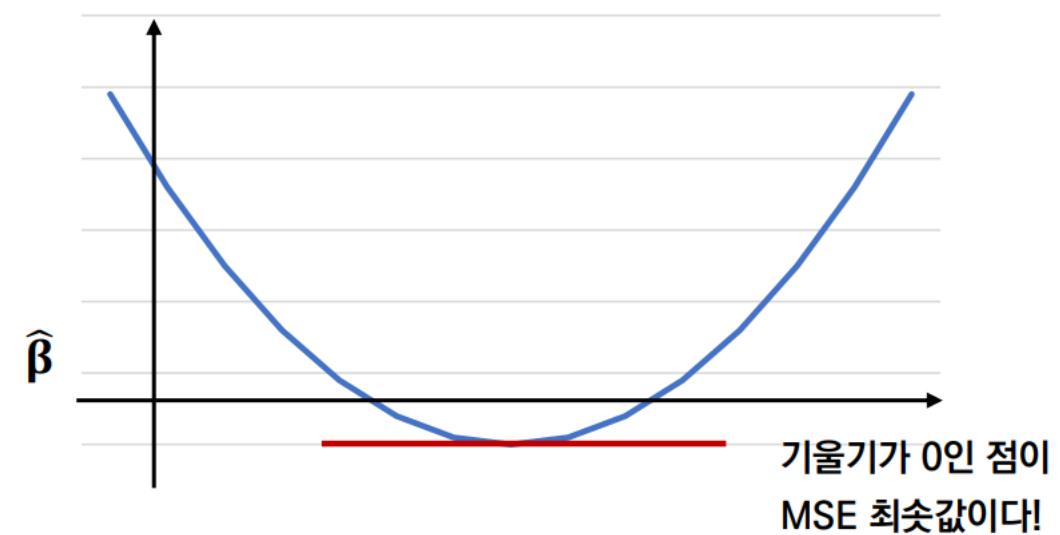
$$MSE = (y - X\hat{\beta})^2 = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$\frac{\partial MSE}{\partial \hat{\beta}} = -2X^T(y - X\hat{\beta})$ 가 0이 되도록 하는 $\hat{\beta}$ 를 찾는다.

$$-X^T(y - X\hat{\beta}) = -X^Ty + X^TX\hat{\beta} = 0$$

$$\rightarrow X^TX\hat{\beta} = X^Ty \rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\therefore \hat{\beta} = (X^TX)^{-1}X^Ty$$



다중선형회귀 : Scaling

데이터 스케일링(Data Scaling) : 서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업.

값을 조정하는 과정이기 때문에 수치형 변수에만 적용해야 함!

>>스케일링 과정을 거쳐, 변수 간의 중요도를 보다 정확하게 파악할 수 있음.

StandardScaling(표준화)

서로 다른 변수의 단위를 없애주는
scaling 기법

$$z = \frac{x_i - \text{mean}(x)}{\text{stddev}(x)}$$

변수 각각의
평균을 0, 분산을 1로 변환

MinMaxScaling(정규화)

서로 다른 변수의 크기를 통일하기
위해 크기를 변환해주는 scaling 기법

$$\frac{x_i - \min(x)}{\max(x) - \min(x)}$$

변수 각각의 값을
모두 0에서 1 사이로 변환

RobustScaling

데이터의 중앙값 = 0, IQR = 1이
되도록 하는 scaling 기법

$$\frac{x_i - \text{median}(x)}{Q3 - Q1}$$

Q3 : x의 제3사분위수
Q1 : x의 제1사분위수

모든 변수들이 같은 scale을 갖게 됨
스케일링 결과가 상대적으로 더
넓은 범위로 분포하게 됨

다중선형회귀 : Scaling(예시)

원본 데이터(head)

Age	22.00	38.0000	26.000	35.0	35.00
Fare	7.25	71.2833	7.925	53.1	8.05

StandardScaling(표준화)

	Age	Fare
0	-0.530377	-0.502445
1	0.571831	0.786845
2	-0.254825	-0.488854
3	0.365167	0.420730
4	0.365167	-0.486337

변환된 데이터

MinMaxScaling(정규화)

	Age	Fare
0	0.271174	0.014151
1	0.472229	0.139136
2	0.321438	0.015469
3	0.434531	0.103644
4	0.434531	0.015713

RobustScaling

	Age	Fare
0	-0.335664	-0.312011
1	0.559441	2.461242
2	-0.111888	-0.282777
3	0.391608	1.673732
4	0.391608	-0.277363

스케일링 결과 확인

train_standard.mean()

Age 2.388379e-16
Fare 3.987333e-18

train_standard.var()

Age 1.001403
Fare 1.001124

train_minmax.min()

Age 0.0
Fare 0.0

train_minmax.max()

Age 1.0
Fare 1.0

train_robust.median()

Age 0.0
Fare 0.0

train_robust_IQR =
train_robust.quantile(0.75)
- train_robust.quantile(0.25)

Age 1.0
Fare 1.0

다중선형회귀 : 선형모형에서의 가설 검정

선형모형에서의 가설

1. 예측 변수들과 연관된 모든 회귀계수들이 0이다.

Ex) $H_0: \beta_1 = \beta_2 = 0$ vs $H_1: \beta_1 \neq 0$ or $\beta_2 \neq 0$

2. 회귀계수들 중 일부분이 0이다.

Ex) $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

3. 회귀계수들 중 일부분이 서로 같은 값을 가진다.

Ex) $H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$

4. 회귀계수들이 특정한 제약조건을 만족한다.

Ex) $H_0: \beta_1 + \beta_2 = 1$ vs $H_1: \beta_1 + \beta_2 \neq 1$

귀무가설 vs 대립가설

귀무가설(H_0): 처음부터 버릴 것을 예상하는 가설.

대립가설(H_1): 귀무가설과 대립하는 가설. 귀무가설처럼 검정을 직접 수행하기는 불가능하며 귀무가설을 기각함으로써 받아들여 채택됨

축소모형 vs 완전모형

완전모형: 기본 다중선형회귀모형 $\rightarrow Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$

축소모형: 완전모형의 부분집합. 귀무가설을 완전모형에 대입한 것!

1. 예측변수들과 연관된 모든 회귀계수들이 0이다. \rightarrow 예) $Y = \beta_0 + \varepsilon$

2. 회귀계수들 중 일부분이 0이다. \rightarrow 예) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

가설 검정 = 축소모형(귀무가설)과 완전모형(대립가설) 중 무엇이 적절?

다중선형회귀 : 선형모형에서의 가설 검정

선형모형에서의 가설 검정 방법

검정통계량 : F-검정이 사용됨.

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

SSE(FM) : 완전모형의 잔차제곱합

SSE(RM) : 축소모형의 잔차제곱합

n-p-1 : 완전모형의 자유도

p+1-k : 축소모형의 자유도 (n-k) – 완전모형의 자유도

(n : 표본의 크기, p : 독립변수의 수, k: 추정될 독립변수의 수)

귀무가설 기각 여부

$F \geq F(p+1-k, n-p-1, \alpha)$ 또는 $p(F) \leq \alpha \rightarrow$ 귀무가설 기각!

즉, F값이 기각치보다 크거나 p값이 유의수준보다 작으면 H_0 기각

예시 : 가설 1

1. 예측 변수들과 연관된 모든 회귀계수들이 0이다.

RM : $H_0 : Y = \beta_0 + \varepsilon$

FM : $H_1 : Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$

$$\text{검정통계량} = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

$$= \frac{[SST - SSE]/p}{SSE/(n-p-1)}$$

$$= \frac{SSR/p}{SSE/(n-p-1)}$$

$$\frac{\text{MSR} \text{ (평균회귀제곱합)}}{\text{MSE} \text{ (평균오차제곱합)}}$$



이 F-statistic을 임계치와 비교!

다중선형회귀 : 다중상관계수 ($R = \sqrt{R^2}$)

다중상관계수

$$Cor(Y, \hat{Y}) = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$$

혹은

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \{ Cor(Y, \hat{Y}) \}^2$$

- 다중회귀분석에서의 반응값(Y)과 예측값(Y hat) 간의 상관계수
- 다중선형회귀 모델이 Y를 얼마나 잘 예측할 수 있는지에 대한 척도
- $-1 \sim 1$ 사이의 값
- 값이 클수록 종속변수의 변동·패턴을 모델이 잘 설명함.

다중선형회귀 : regression assumptions

Regression assumptions(OLS)

등분산성

잔차의 분산은 모든 설명변수들에 대해 일정한 분산을 가짐

선형성

설명변수와 종속변수 간의 관계는 항상 선형적으로 설명됨.
즉, 종속변수는 설명변수들의 1차항의 결합으로 표현될 수 있음

잔차의 정규분포

잔차들의 분포는 정규분포를 따라야 함

잔차의 독립성

서로 다른 설명변수에 대한 잔차들은 서로 독립적임

다중공선성 존재 X

설명변수 사이에 다중공선성이 존재하지 않음

다중공선성?

정의

회귀분석에서 사용된 모형의 일부 설명 변수가 다른 설명 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상. 수리적으로는, 어떤 독립변수가 다른 독립변수들과 완벽한 선형 독립이 아닌 경우를 의미.

왜 문제가 되는가?

다중공선성 문제가 발생하면,

- 다중공선성에 해당하는 변수들 각각의 설명력이 약해짐
- 설명력이 약해짐으로써 설명변수들의 표준오차와 p-value가 정상 수치보다 더 커지게 됨
- 즉, 부정확한 회귀 결과가 도출됨

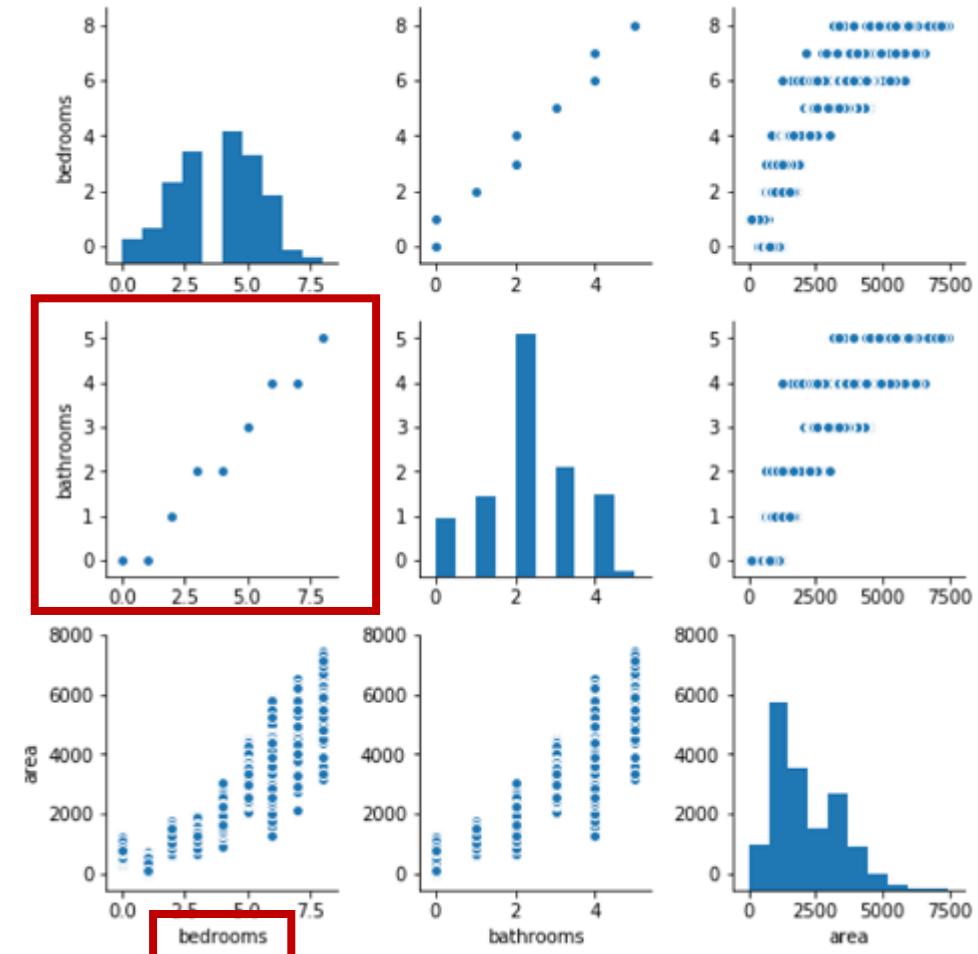
다중공선성 – 예시

데이터

house_id	neighborhood	area	bedrooms	bathrooms	style	price	
0	1112	B	1188	3	2	ranch	598291
1	491	B	3512	5	3	victorian	1744259
2	5952	B	1134	3	2	ranch	571669
3	3525	A	1940	4	2	ranch	493675
4	5108	B	2208	6	4	victorian	1101539

bedrooms, bathrooms, area 간 상관 관계가 있는가?
)) 산점도를 통해 독립 변수간 1대 1 상관 관계 파악 가능
예) 2행 1열은 bathrooms와 bedrooms간의 그래프로,
강한 양(+)의 상관 관계 존재

산점도 : 변수 간의 상관관계 파악



다중공선성 – 예시(Y : Price)

변수를 DROP하지 않은 경우

	coef	std err	t	P> t	[0.025	0.975]
intercept	1.007e+04	1.04e+04	0.972	0.331	-1.02e+04	3.04e+04
bedrooms	-2925.8063	1.03e+04	-0.285	0.775	-2.3e+04	1.72e+04
bathrooms	7345.3917	1.43e+04	0.515	0.607	-2.06e+04	3.53e+04
area	345.9110	7.227	47.863	0.000	331.743	360.079



단순선형회귀

	coef	std err	t	P> t	[0.025	0.975]
intercept	-9.485e+04	1.08e+04	-8.762	0.000	-1.16e+05	-7.36e+04
bedrooms	2.284e+05	2646.744	86.289	0.000	2.23e+05	2.34e+05

Bedrooms의 계수 : -2925.8063
침실 개수와 집 값 : 음(-)의 상관 관계

Bedrooms의 계수 : 2.284e+05
침실 개수와 집 값 : 양(+)의 상관 관계

) 침실 개수의 집값은 양의 상관 관계를 보이는 것이 맞음.

변수를 Drop하지 않은 다중선형회귀에서 두 변수 간의 관계가 음수로 나온 것은, 다중공선성으로 인해 회귀 결과가 왜곡된 것.

즉, 다중공선성 문제를 가지는 변수들은 굉장히 불안정한 계수값을 보여주게 됨

∴ 다중공선성 반드시 제거!

다중공선성 판별 :VIF score 활용

VIF?

VIF : 각 독립 변수를 종속 변수로 보고, 나머지 독립 변수들을 이용하여 선형 회귀 모델을 적합시킨 후, 그 설명력을 이용하여 계산

$$VIF_i = \frac{1}{1 - R_i^2} \quad \rightarrow \quad \text{Var}(\hat{\beta}_j^\bullet) = \frac{\sigma^{*2}}{n\text{var}(x_j)} VIF_j$$

(단, 여기서의 R-squared는 모델 전체의 설명력이 아니라, 특정 독립변수 i를 종속변수로 두고 나머지 독립변수들을 독립변수로 하여 진행한 회귀분석 모델의 설명력)

특정 독립변수가 다른 독립변수들에 의해 설명되는 정도가 높음 >> 독립변수에 해당하는 R-squared 값이 커짐
>> VIF 값이 커짐 >> VIF 값이 커지면 해당 독립변수의 분산 값도 커지게 됨!

VIF Score를 통한 다중공선성 문제 판별

VIF Score ≥ 5 : 독립변수가 다중공선성 문제를 가짐

VIF Score ≥ 10 : 해당 독립변수를 제외할 것

단, 판단 역치는 무조건적으로 고정된 것은 아니고, 데이터의 특성에 따라 변동 가능함

	Variable	VIF Factor
0	host_response_time	5.712469
1	host_listings_count	1.261886
2	host_since_date	4.678013
3	host_verifications_counts	21.857386
4	host_response_rate(d)	34.148903
5	host_acceptance_rate(d)	14.537609

다중선형회귀분석 : 변수 선택법

정의

종속변수 Y 가 있을 때, 이 Y 에 영향을 미칠 것으로 예상되는 설명변수 X 를 선택하는 방법

1. 모든 가능한 회귀(all possible regression)
2. 전진선택법(forward selection)
3. 후진제거법(backward elimination)
4. 단계별 회귀방법(stepwise regression)

주의할 점

각각의 방법을 통해 선택된 변수가 다를 수도 있음.

∴ 주어진 데이터의 성격에 따라 적절한 변수 선택법을 선택하여 회귀분석을 진행할 것!

(즉, 이들 중 어느 것이 최적 회귀방정식임을 보장할 수 없고, 또 최적 회귀방정식이 둘 이상인 경우도 있음!)

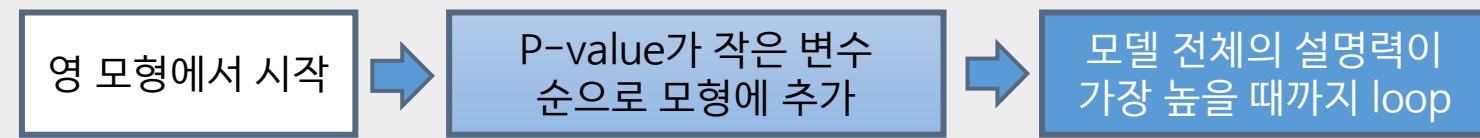
다중선형회귀분석 : 변수 선택법

모든 가능한 회귀

모든 가능한 변수들의 조합을 모형화하여, 그 중에서 최적의 모형을 선택

전진선택법

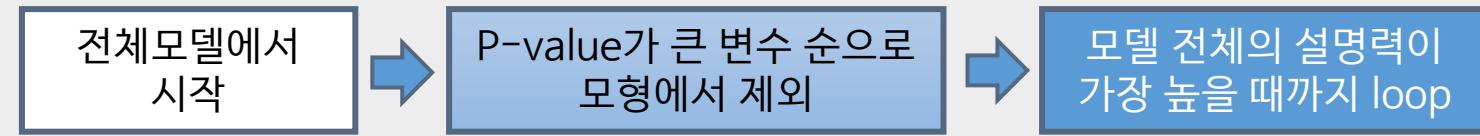
독립변수들 중 가장 큰 영향을 미칠 것으로 판단되는 변수부터 (유의성이 높은 변수부터) 하나씩 선택하여,
더 이상 중요한 변수가 없다고 판단될 때 변수의 선택을 중단.



AIC
BIC
고려!

후진제거법

독립변수들 중 가장 작게 영향을 미칠 것으로 판단되는 변수부터(유의성이 작은 변수부터) 하나씩 제거.

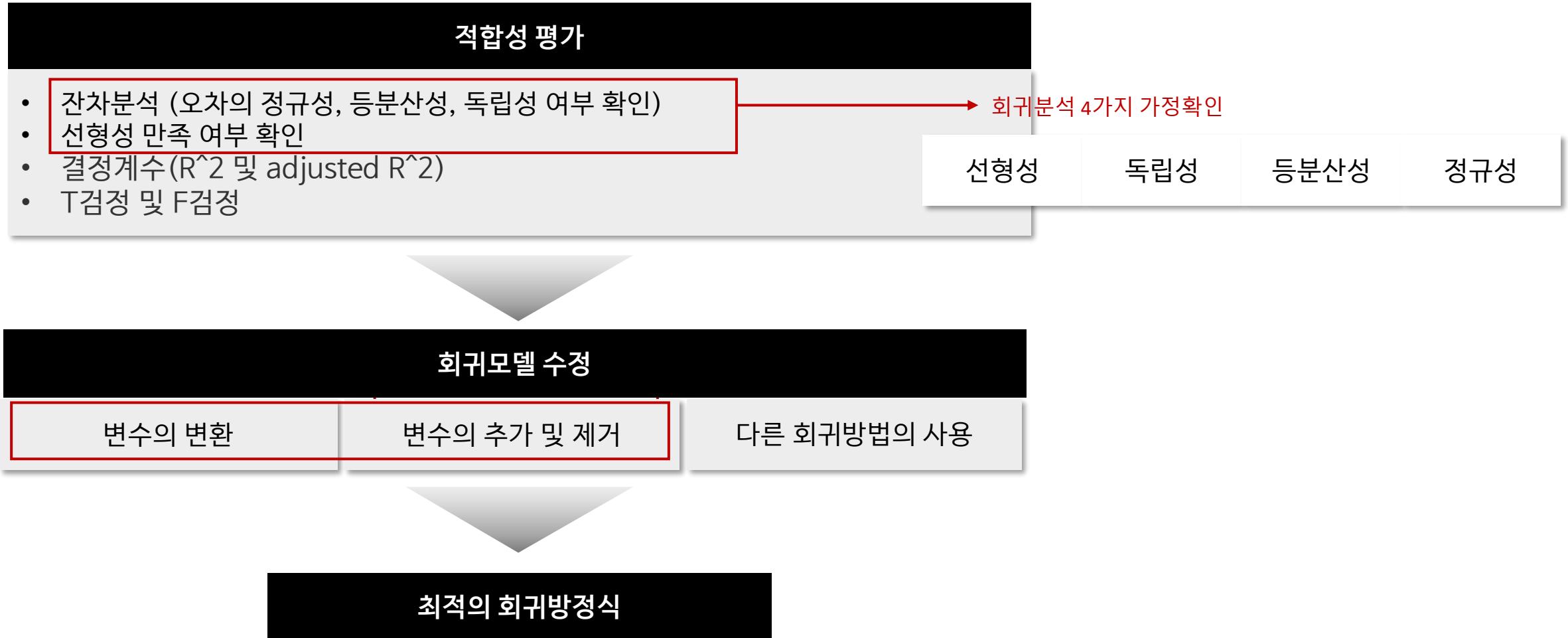


단계별 회귀방법

종속변수의 추가와 제거를 반복

Multiple Regression Model – Part 2

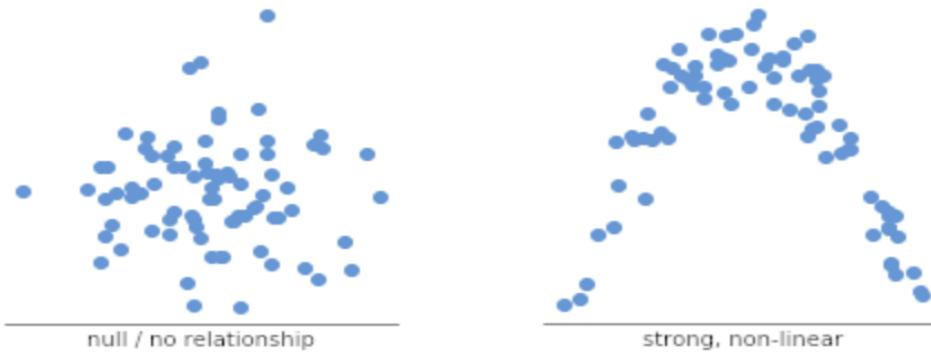
회귀모델 적합성 평가



선형성

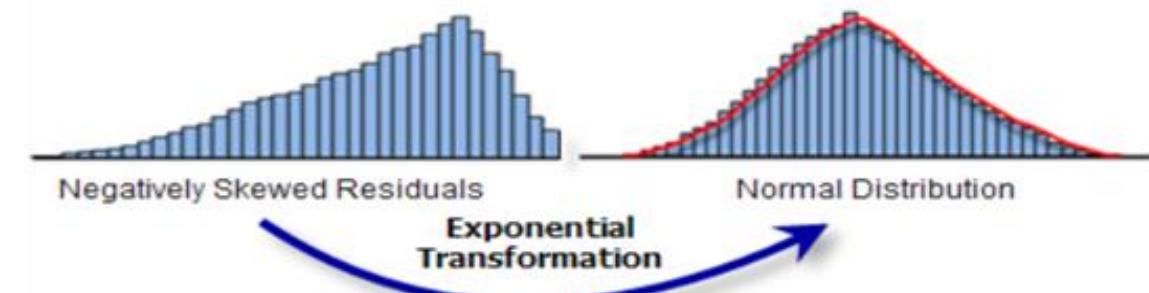
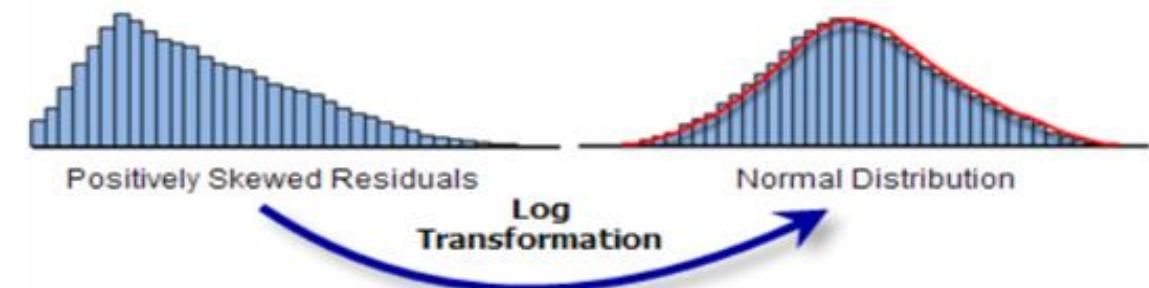
독립변수와 종속변수 간의 선형 관계 존재

› 선형성 위배 예시



› 선형성 위배 해결방안

변수 변환(로그, 제곱근 등)



› 선형성이 위배될 경우

비선형적인 데이터의 경우 변동성을 설명하는데 한계점이 존재하며, 회귀모델의 예측력 저하로 이어질 우려가 있음.
(해당 가정의 위배는 R^2 의 저하로 관측 가능함.)

독립성

더빈-왓슨 검정(Durbin-Watson test)

- 시계열 데이터에서 오차항의 자기상관을 검정하기 위한 통계적 방법

› 주로 회귀분석에서 잔차의 독립성을 확인하는 데 사용

OLS Regression Results						
Dep. Variable:	dist	R-squared:	0.651			
Model:	OLS	Adj. R-squared:	0.644			
Method:	Least Squares	F-statistic:	89.57			
Date:	Sat, 19 Dec 2020	Prob (F-statistic):	1.49e-12			
Time:	18:40:59	Log-Likelihood:	-206.58			
No. Observations:	50	AIC:	417.2			
Df Residuals:	48	BIC:	421.0			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
Intercept	-17.5791	6.758	-2.601	0.012	-31.168	-3.990
speed	3.9324	0.416	9.464	0.000	3.097	4.768
Omnibus:	8.975	Durbin-Watson:	1.676			
Prob(Omnibus):	0.011	Jarque-Bera (JB):	8.189			
Skew:	0.885	Prob(JB):	0.0167			
Kurtosis:	3.893	Cond. No.	50.7			

› 더빈-왓슨 검정통계량

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2} = \frac{\sum_{t=2}^n e_t^2 - 2e_t e_{t-1} + e_{t-1}^2}{\sum_{t=2}^n e_t^2} \approx 2(1 - \hat{\rho})$$

이 때 $\hat{\rho}$ 는 잔차 간 자기상관계수로, -1과 1의 값을 가진다.

- 상관관계가 양일 경우: $\hat{\rho}$ 가 1에 가까워지고, d 값은 0에 가까움.
- 상관관계가 음일 경우: $\hat{\rho}$ 가 -1에 가까워질수록 d 값은 4에 가까움.

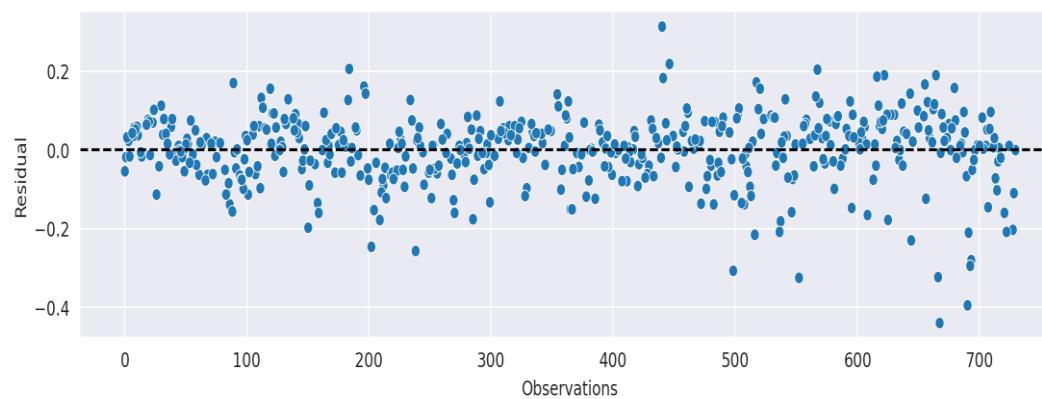
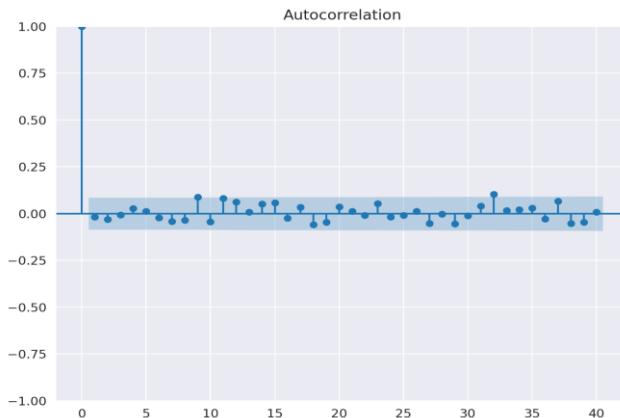
› 검정통계량에 따른 해석



* 미정의 경우 잔차 간 독립성에 대해 결론을 내릴 수 없음.

독립성

› ACF(위), Residual Plot(아래)



› 독립성이 위배될 경우(자기상관)

모형 예측력 감소

Spurious regression(가성회귀) 문제 발생

: P-value가 쉽게 낮아지게 되어 무의미한 변수들도 유의미하게 나타날 가능성이 높아짐

› 독립성 위배 해결방안

시계열 모델 사용

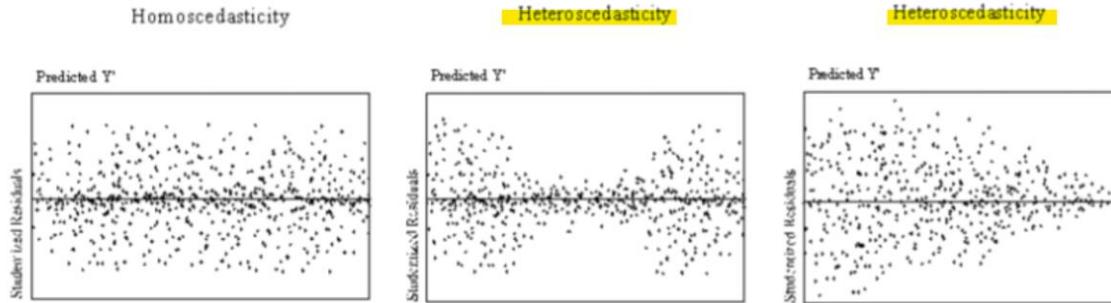
데이터 변환(차분이나 변환)

외부변수 추가

비선형 모델

등분산성

회귀모형을 통해 예측된 값이 크던 작던 관계없이, 모든 값들에 대해 잔차의 분산이 동일하다는 가정



- 위 그래프는 예측값(가로축)에 따라 잔차가 어떻게 달라지는지 보여줌
- 왼쪽의 그래프 잔차는 고른 분포!
- 가운데와 오른쪽 그래프는 잔차가 고르지 않고 편재되어있음

```
ncvTest(mfit)  (p = 0.18632 > 0.05이므로 등분산성 만족)
# Non-constant Variance Score Test
# Variance formula: ~ fitted.values
# Chisquare = 1.746514, Df = 1, p = 0.18632
```

› 등분산성이 위배될 경우 (이분산성)

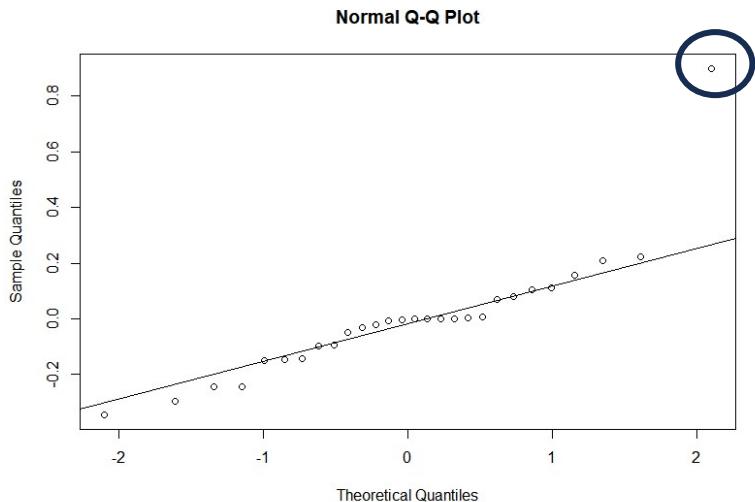
독립 변수가 변할 때 종속 변수의 분산도 변하는 문제가 발생

› 등분산성 위배 해결방안

변수 변환(로그,제곱근 등)

정규성

잔차가 정규분포를 따른다는 가정, 잔차가 정규분포를 띠면 점들이 점선을 따라 배치



〉 정규성이 위배될 경우

오차항이 비정규분포인 경우 회귀 모수에 대한 유의성 검정 및 신뢰구간 추정이 유효하지 못한 문제 발생

〉 정규성 위배 해결방안

데이터변환(로그, 제곱근)

이상치 제거

일반화 선형모델
(glm: 다른 오차 분포 가정)

로버스트 회귀

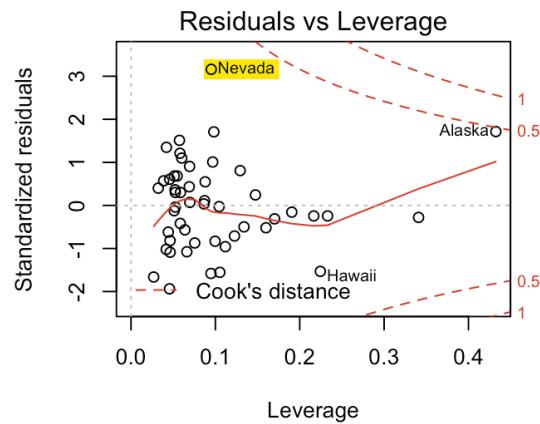
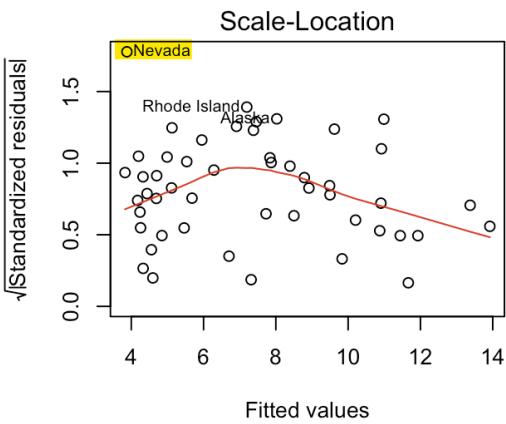
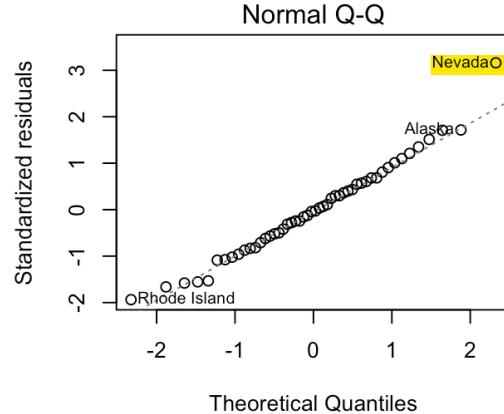
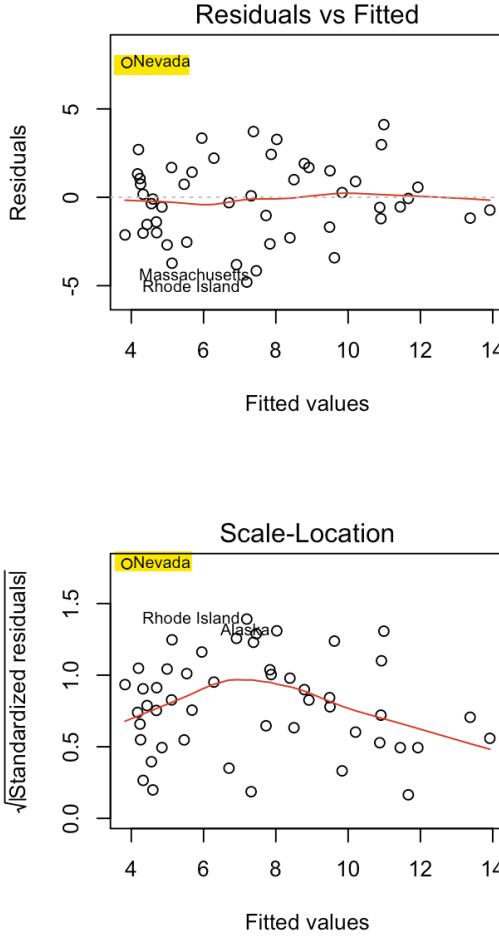
변수 선택

변수 축소

```
install.packages("car", dependencies = T)
library(car)

shapiro.test(residuals(mfit))    (p-value< 0.05 이상이면 정규성 만족)
# Shapiro-Wilk normality test
#
# data:  residuals(mfit)
# W = 0.98264, p-value = 0.6672
```

모델 평가와 해석



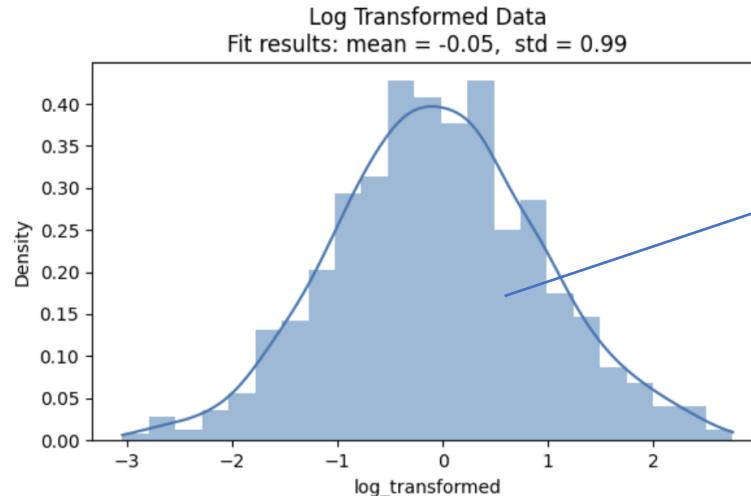
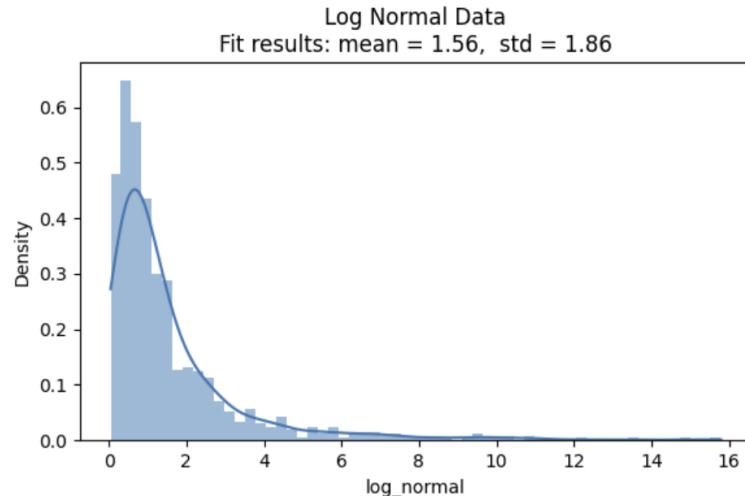
Q-Q plot에서 Nevada 주의 데이터가 이상치인 것을 제외하고 통계의 가정을 잘 만족시키는 것을 알 수 있음.

〉〉 이상치(outlier)

: 회귀모형으로 잘 예측되지 않는 관측치(아주 큰 양수/음수의 residual)
정상의 범주(전체적 패턴)에서 벗어난 값

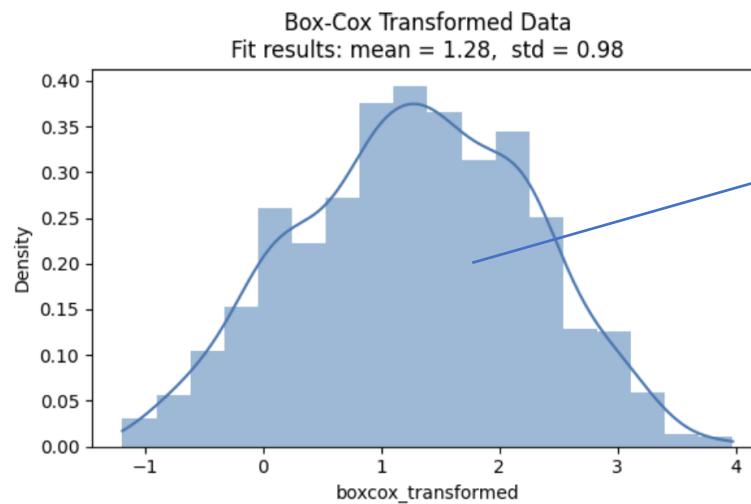
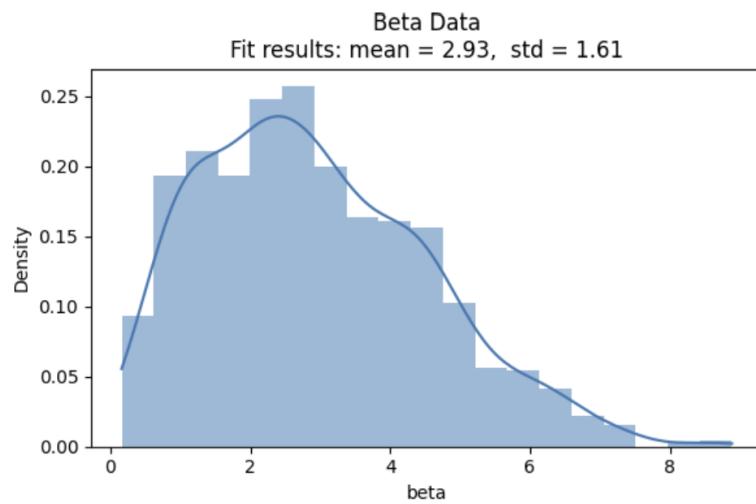
분석결과의 왜곡 발생 가능성 존재,
분석정확도 감소

회귀가정 위반시 해결책



» Log 변환

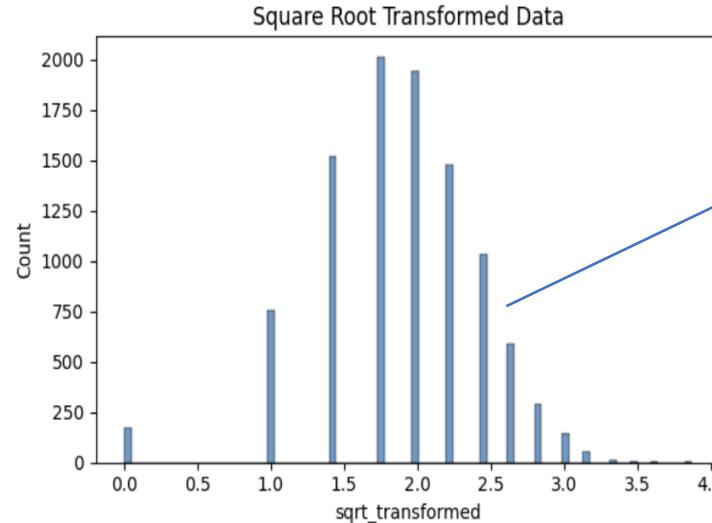
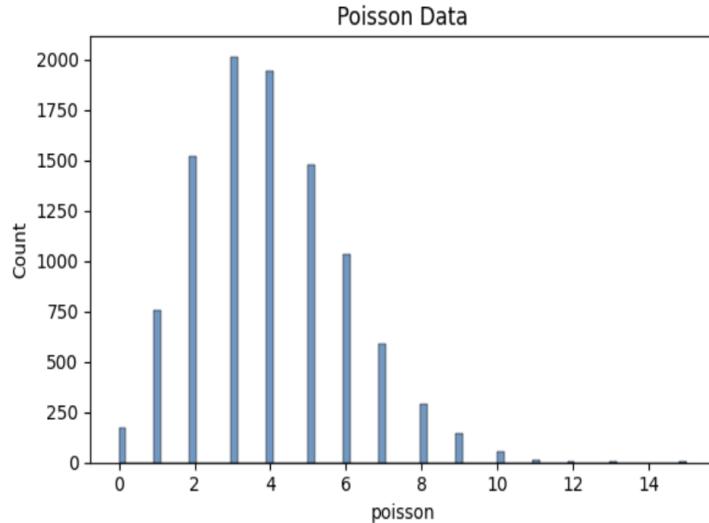
변수 값에 자연로그를 취해주는 방식
0에 가까운 값들이 모여있는 입력값을
넓은 범위로 펼칠 수 있음



» Box-Cox 변환

-2부터 2까지의 범위를 가지며, 최적값은
에러의 제곱합을 최소화함.
Box-cox의 경우, 원래 데이터가 양의 값을
가질 때만 가능

회귀가정 위반시 해결책



»제곱근 변환

변수 값에 제곱근을 적용
오른쪽으로 꼬리가 긴 데이터를 왼쪽으로
모아 대칭화에 유용
로그 변환에 비해 변환된 크기가 작음

변수변환의 이유

- › 치우친 분포를 가지는 변수의 경우 모델 결과 예측 정확도가 떨어지므로
- › 첨도와 왜도가 줄어들며 정규성 향상에 기여
- › 큰 값들 및 이상치에 의해 왜곡된 상관관계 해결 가능

회귀모델 선택법

모델 성능지표: R^2 및 Adjusted R^2, MSE

R^2

- 독립변수가 종속변수 변동의 몇 %인지를 설명하는 지표
- 독립변수의 유의성 여부와 관계없이 독립변수의 수가 많아지면 결정계수가 높아지는 단점이 있음

Adjusted R^2

- 결정계수의 단점을 보완하기 위해 수정된 결정계수 (R_a^2 : adjusted R^2) 활용
- 수정된 결정계수는 결정계수보다 작은 값으로 산출되는 특징이 있음

$$R_{adj}^2 = 1 - \left[\frac{n-1}{n-(p+1)} \right] \frac{SSE}{SST} \leq 1 - \frac{SSR}{SST} = R^2$$



변수의 개수에 대한 penalty를 추가해

유의하지 않은 변수가 추가될 경우 결정계수가 증가하지 않도록 함

MSE

- 실제값에서 예측값을 뺀 값을 제곱한 SSE를 SSE의 자유도인 n으로 나눈 값
- 값이 작을수록 좋으나 과하게 줄이면 과적합의 가능성 있음
- 특이값에 민감함: 특이값 존재시 수치 많이 증가함

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

회귀모델 선택법

모델 성능지표: AIC 및 BIC

AIC

- 변수 선택 방법은 일반적으로 AIC를 기반으로 하여 모델의 품질을 평가함
- AIC값이 낮을수록 더 우수한 모델

BIC

- 샘플 크기에 대한 추가적인 패널티를 부여해 큰 데이터 세트에서 과적합을 방지하는데 도움
- 표본의 크기가 커지면 AIC보다 BIC가 더 잘 맞음
- BIC값이 낮을수록 더 우수한 모델

$$AIC = -2(\text{log-likelihood}) + 2k$$

- Log-likelihood: 로그 가능성. 가능성은 확률 분포의 모수가 어떤 확률변수의 표집값과 일관되는 정도
- 2k: 모형에 추가된 변수의 개수

$$BIC = -2(\text{log-likelihood}) + k\log(n)$$

- Log-likelihood: 로그 가능성. 가능성은 확률 분포의 모수가 어떤 확률변수의 표집값과 일관되는 정도
- n: 표본의 크기

회귀모델 선택법

Stepwise regression을 활용한 변수선택법

Curse of dimensionality

차원의 저주: 독립변수의 수가 너무 많아지면
오히려 그 성능은 낮아지는 현상

Stepwise regression

각 단계마다 영향력을 주는 변수들을
제거하거나 더하면서 변수를 선별함

Optimal regression equation

최적의 회귀방정식

Stepwise regression

- 모든 경우의 수를 파악하는 것보다 상관성이 높은 회귀계수를 먼저 찾아내는 법
- 단계별로 영향력을 주는 변수들을 확인 가능하다는 장점이 있음

전진선택법

절편만 있는 상수 모델에서 시작해 가장 중요한 설명변수부터 차례대로 추가

후진제거법

모든 후보 설명변수가 포함된 모델에서 시작해 가장 영향이 적은 변수부터 제거

단계적방법

추가& 제거 반복

기준 통계치를
개선시키면 추가,
아니면 추가를 멈춤

더 이상 유의하지 않은
변수가 없을 때까지
제거

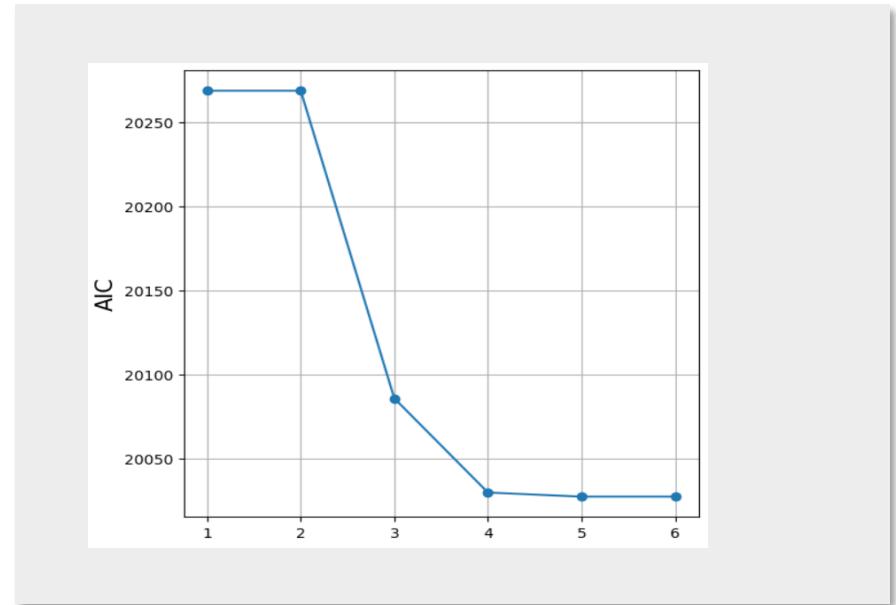
“최적의 회귀방정식”을 찾기 위한 과정

전진선택법

```
while len(variables) > 0:
    remainder = list(set(variables) - set(forward_variables))
    pval = pd.Series(index=remainder) ## 변수의 p-value
    ## 기준에 포함된 변수와 새로운 변수 하나씩 돌아가면서
    ## 선형 모형을 적합한다.
    for col in remainder:
        X = X_train[forward_variables+col]
        X = sm.add_constant(X)
        model = sm.OLS(y,X).fit(disp=0)
        pval[col] = model.pvalues[col]

    min_pval = pval.min()
    if min_pval < sl_enter: ## 최소 p-value 값이 기준 값보다 작으면 포함
        forward_variables.append(pval.idxmin())
        ## 선택된 변수들에 대해서
        ## 어떤 변수를 제거할지 고른다.
        while len(forward_variables) > 0:
            selected_X = X_train[forward_variables]
            selected_X = sm.add_constant(selected_X)
            selected_pval = sm.OLS(y,selected_X).fit(disp=0).pvalues[1:] ## 절편항의 p-value는 뺀다
            max_pval = selected_pval.max()
            if max_pval >= sl_remove: ## 최대 p-value값이 기준값보다 크거나 같으면 제외
                remove_variable = selected_pval.idxmax()
                forward_variables.remove(remove_variable)
            else:
                break
        step += 1
        steps.append(step)
        adj_r_squared = sm.OLS(y,sm.add_constant(X_train[forward_variables])).fit(disp=0).rsquared_adj
        adj_r_squared_list.append(adj_r_squared)
        sv_per_step.append(forward_variables.copy())
    else:
        break
```

>AIC 변화



도요타자동차가격데이터

<https://www.kaggle.com/klkwak/toyotacorollacsv>

후진제거법

```
def backward_regression(X, y,
                       initial_list=[],
                       threshold_out = 0.05, # P-value 임계값 (제거 기준)
                       feature_list = X_train.columns.tolist()
                      ):

    sv_per_step = [] ## 각 스텝별로 선택된 변수들
    AIC_list = [] ## 각 스텝별 수정된 결정계수
    steps = [] ## 스텝
    step = 0
    included = feature_list
    while True:
        changed=False
        model = sm.OLS(y, sm.add_constant(pd.DataFrame(X[included]))).fit(disp=0)
        # use all coefs except intercept
        pvalues = model.pvalues.iloc[1:] # 각 feature의 P값을 의미함
        worst_pval = pvalues.max() # P 값이 가장 높은 것 설정
        if worst_pval > threshold_out:
            changed=True
            worst_feature = pvalues.idxmax()
            included.remove(worst_feature)

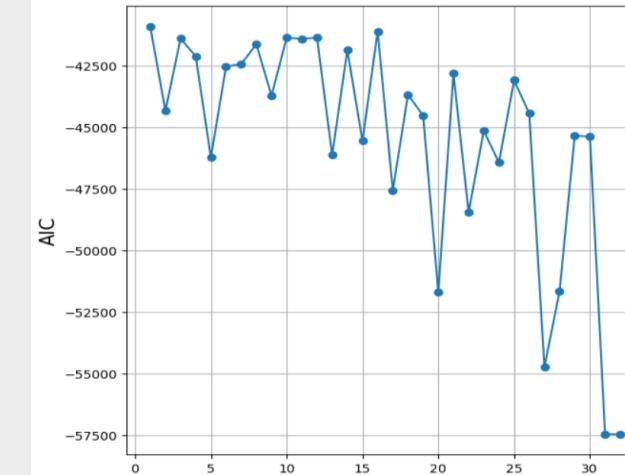
        step += 1
        steps.append(step)
        AIC = sm.OLS(y, sm.add_constant(pd.DataFrame(X[included]))).fit(disp=0).aic
        AIC_list.append(AIC)
        sv_per_step.append(included.copy())

        if not changed:
            break

    return included,step,steps,AIC_list,sv_per_step

backward_valriables_function,step,steps,AIC_list,sv_per_step = backward_regression(X_train, y_train)
```

>AIC 변화



전진선택법

```
minModel <- lm(Murder ~ 1, data = states)
fwd_model <- step(minModel, direction = "forward",
                   scope = (Murder ~ Population + Illiteracy + Income + Frost), trace = 1)
# Start: AIC=131.59 # 독립변수 추가 전 AIC
# Murder ~ 1 # 독립변수 없음
#
#           Df Sum of Sq   RSS   AIC
# + Illiteracy  1     329.98 337.76 99.516
# + Frost      1     193.91 473.84 116.442
# + Population 1     78.85 588.89 127.311
# + Income     1     35.35 632.40 130.875
# <none>          667.75 131.594
#
# Step: AIC=99.52 # 가장 낮은 AIC의 독립변수 Illiteracy 추가 후 AIC (낮아짐)
# Murder ~ Illiteracy # 독립변수 Illiteracy 추가
#
#           Df Sum of Sq   RSS   AIC
# + Population  1     48.517 289.25 93.763
# <none>          337.76 99.516
# + Frost      1     5.387 332.38 100.712
# + Income     1     4.916 332.85 100.783
#
# Step: AIC=93.76 # 가장 낮은 AIC의 독립변수 Illiteracy, Population 추가 후 AIC (최종)
# Murder ~ Illiteracy + Population # 독립변수 Illiteracy, Population 추가
#
#           Df Sum of Sq   RSS   AIC
# <none>          289.25 93.763
# + Income    1  0.057022 289.19 95.753
# + Frost     1  0.021447 289.22 95.759
```

절편만 있는 상수 모델에서 시작해 가장 중요한 설명변수부터 차례대로 추가함

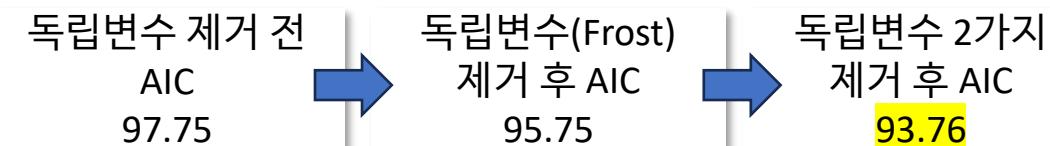


AIC는 계속 감소하여 93.76에 이를 성능의 증가

후진제거법

```
fullModel <- lm(Murder ~ ., data = states) # 모든 독립변수를 모델링하고 backward
reduce_model <- step(fullModel, direction = "backward")
# Start: AIC=97.75 # 모든 독립변수를 모델링 했을 때의 AIC
# Murder ~ Population + Illiteracy + Income + Frost # 모든 독립변수
#
#          Df Sum of Sq    RSS     AIC
# - Frost      1   0.021 289.19  95.753
# - Income      1   0.057 289.22  95.759
# <none>           289.17  97.749
# - Population  1   39.238 328.41 102.111
# - Illiteracy   1   144.264 433.43 115.986
#
# Step: AIC=95.75 # Frost를 제거 후 모델링 했을 때의 AIC (낮아짐)
# Murder ~ Population + Illiteracy + Income # 제일 AIC 기여도가 낮은 Frost를 제거한 나머지 독립변수
#
#          Df Sum of Sq    RSS     AIC
# - Income      1   0.057 289.25  93.763
# <none>           289.19  95.753 # Frost 제거
# - Population  1   43.658 332.85 100.783
# - Illiteracy   1   236.196 525.38 123.605
#
# Step: AIC=93.76 # Frost, Income를 제거 후 모델링 했을 때의 AIC (최종)
# Murder ~ Population + Illiteracy # 그다음 AIC 기여도가 낮은 Income를 제거한 나머지 독립변수
#
#          Df Sum of Sq    RSS     AIC
# <none>           289.25  93.763 # Income 제거
# - Population  1   48.517 337.76  99.516
# - Illiteracy   1   299.646 588.89 127.311
```

모든 변수를 넣고 AIC기여도가 가장 낮은 것부터 제거함



AIC는 계속 감소하여 93.76에 이를 성능의 증가

Pre 실습 세션

독립변수의 형태

Income	Age	Marital Status
\$45,000	23	Single
\$48,000	25	Single
\$54,000	24	Single
\$57,000	29	Single
\$65,000	38	Married
\$69,000	36	Single
\$78,000	40	Married
\$83,000	59	Divorced
\$98,000	56	Divorced
\$104,000	64	Married
\$107,000	53	Married



Income	Age	Married	Divorced
\$45,000	23	0	0
\$48,000	25	0	0
\$54,000	24	0	0
\$57,000	29	0	0
\$65,000	38	1	0
\$69,000	36	0	0
\$78,000	40	1	0
\$83,000	59	0	1
\$98,000	56	0	1
\$104,000	64	1	0
\$107,000	53	1	0

- Situation1. 연속형

X variable: Age, Y variable: Income

$$Income = \beta_0 + \beta_1 Age + \varepsilon$$

- 즉, 회귀식에서 Age 가 1 증가하면, Income 이 $\widehat{\beta}_1$ 증가함.

- Situation2. 범주형

X variable: Marital Status & Age, Y variable: Income

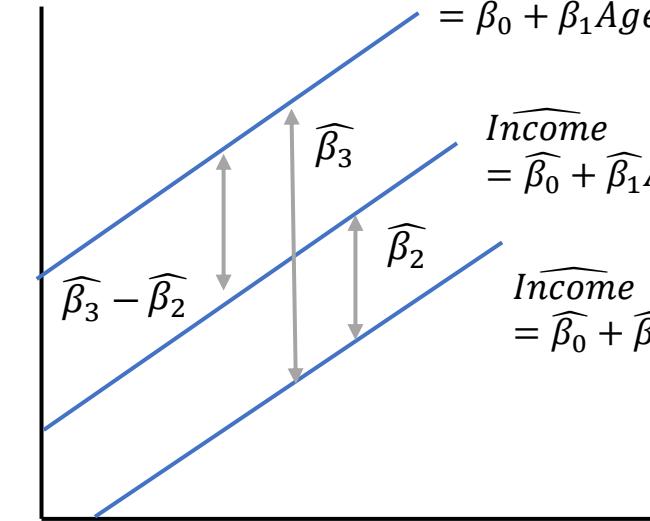
$$Income = \beta_0 + \beta_1 Age + \beta_2 Married + \beta_3 Divorced + \varepsilon$$

- 더 이상 독립변수 증가에 따라, Y가 $\widehat{\beta}_2$ 증가한다고 해석할 수 없음.

$$\widehat{Income} = \widehat{\beta}_0 + \widehat{\beta}_1 Age + \widehat{\beta}_3 (Divorced)$$

$$\widehat{Income} = \widehat{\beta}_0 + \widehat{\beta}_1 Age + \widehat{\beta}_2 (Married)$$

$$\widehat{Income} = \widehat{\beta}_0 + \widehat{\beta}_1 Age (Single)$$



$$Married = \begin{cases} 1 & \text{when married} \\ 0 & \text{o.w} \end{cases}$$

$$Divorced = \begin{cases} 1 & \text{when divorced} \\ 0 & \text{o.w} \end{cases}$$

when Married & Divorced = 0, Single

회귀분석 함수

회귀분석 함수는 1) scikit-learn 2) statsmodels를 이용할 수 있다.

PTRATIO	B	LSTAT	MEDV
19.2	396.90	21.14	19.7
19.2	396.90	14.10	18.3
19.2	396.90	12.92	21.2
19.2	395.77	15.10	17.5
19.2	396.90	14.33	16.8
21.0	391.99	9.67	22.4
21.0	396.90	9.08	20.6
21.0	396.90	5.64	23.9
21.0	393.45	6.48	22.0
21.0	396.90	7.88	11.9

X Y

scikit-learn

```
from sklearn.linear_model import LinearRegression  
  
# LinearRegression fit  
model = LinearRegression()  
model.fit(X_train, y_train)  
  
# Predict  
predict_y = model.predict(X_test)
```

```
# 회귀계수 확인  
model.coef_  
model.intercept_
```

- X, y로 나누어 model fit
- summary 제공하지 않음
(p-value나 R^2 등을 알 수 없음)

statsmodels

```
import statsmodels.formula.api as smf  
  
# LinearRegression fit  
model = smf.ols('MEDV ~ LSTAT + B', data=Boston)  
results = model.fit()
```

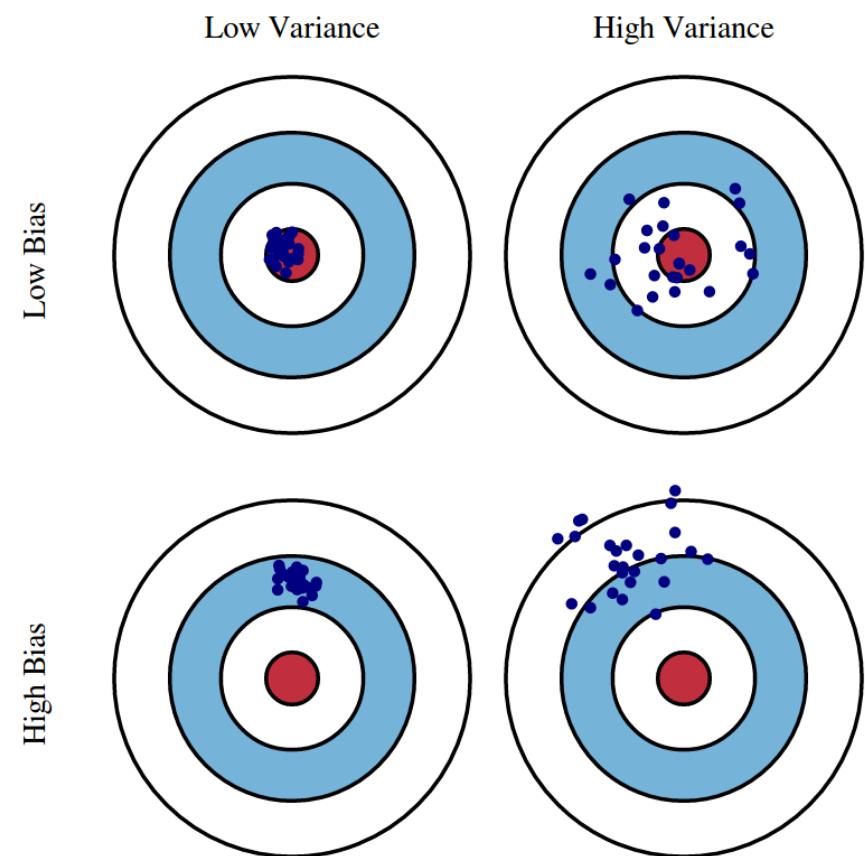
```
# Predict  
predict_y = result.predict(X_new)  
  
results.summary() # coef, p-value, R^2
```

- 데이터 전체 입력해 필요 변수만
(변수 추가 및 제거면에서 편리)
- summary 제공

Ridge & Lasso Regression

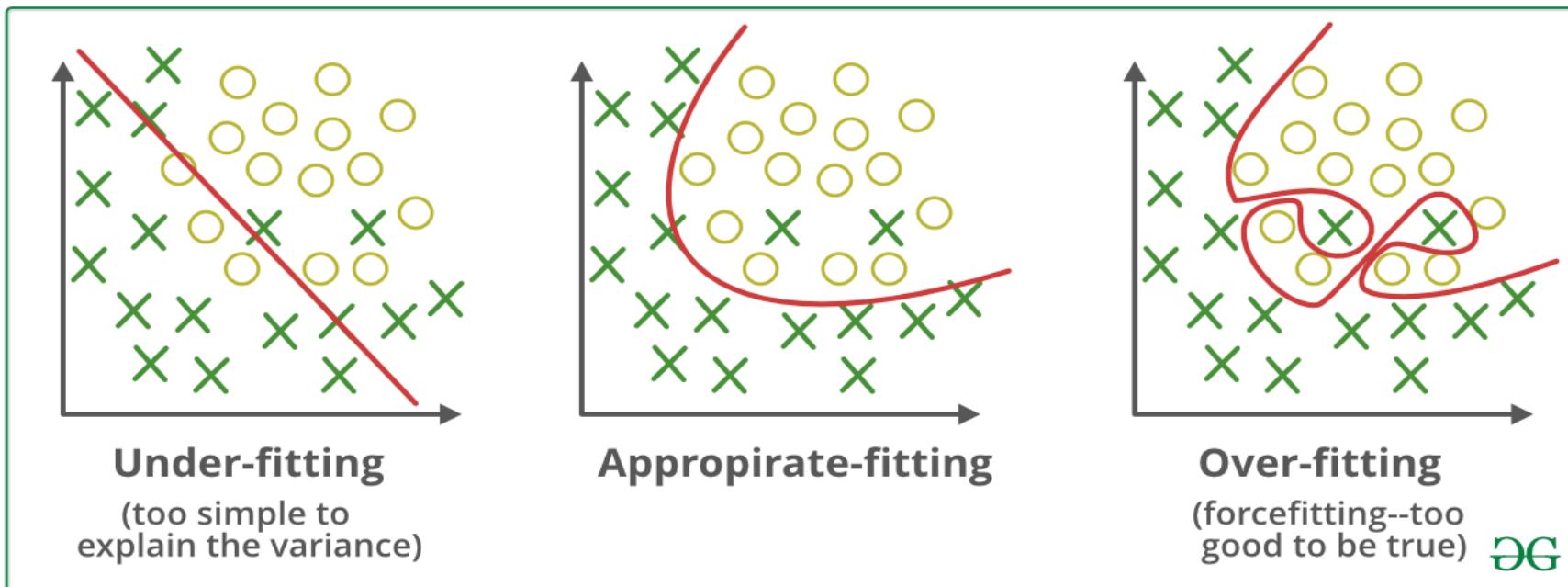
Bias and Variance

- **Bias (편향)** : error because the model cannot represent the concept
- **Variance (분산)** : error because the model overreacts to small changes (noise)
- **Total loss = Bias + Variance**



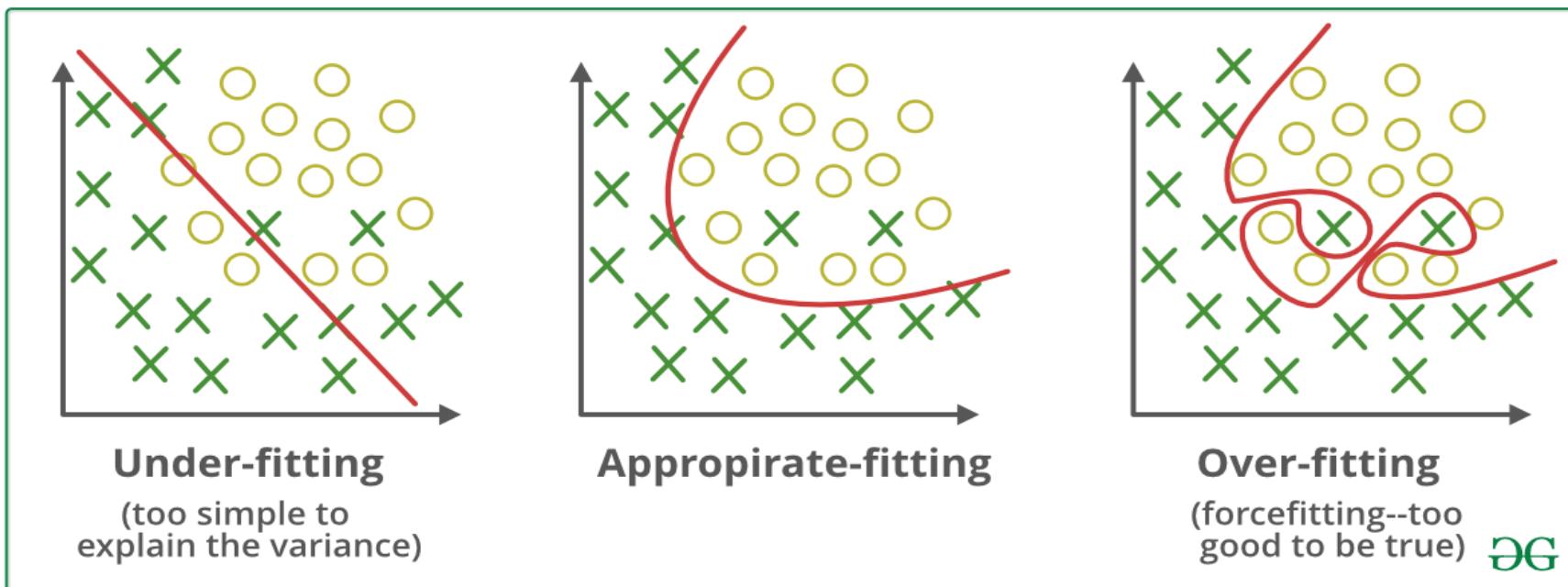
Underfitting

- because of using too simpler model than actual data distribution
- High bias**



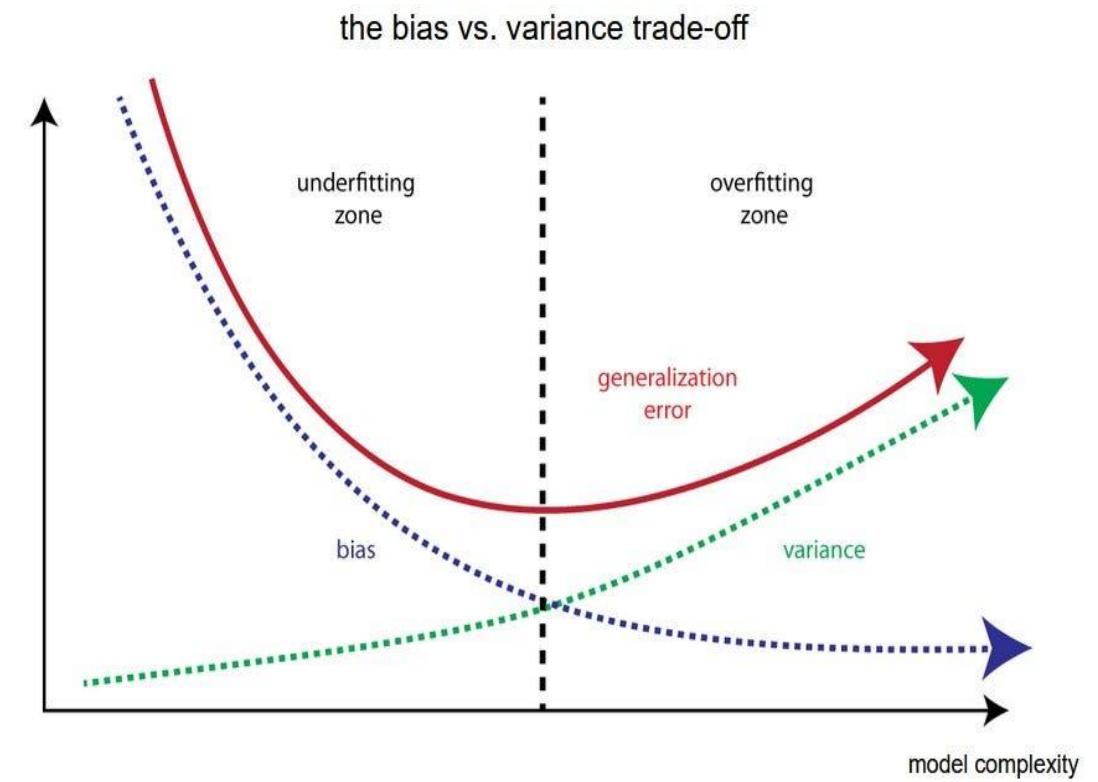
Overfitting

- because of using more complex model than actual data distribution
- High variance



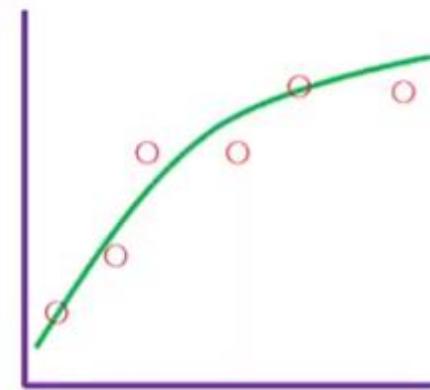
Bias-variance trade-off

- Split into two objectives:
 1. Test error \approx Train error
 2. Train error ≈ 0
- Objective 1: make “Test error \approx Train error”
 - Failure : overfitting \rightarrow high variance and low bias
 - If a model is too complex
- Objective 2 : make “Train error ≈ 0 ”
 - Failure : underfitting \rightarrow high bias and low variance
 - If a model is too simple
- The two objectives have trade-off between approximation and generalization w.r.t model complexity



Regularization concept

- 목적 : overfitting 방지, 일반화 성능 향상
- 일반적으로 회귀 모델에서 계수(β or w)의 크기가 큰 경우, 모델은 overfitting될 가능성이 높아짐.
- 계수의 크기를 제한함으로써, 모델이 예측할 때 사용하는 변수의 개수를 줄임. 즉, 모델의 복잡도를 줄임.



$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

Lasso, Ridge 요약 비교

- 공통점
 - 1. 회귀 모델에서 계수 값을 제한하기 위한 Regularization 방법
 - 2. Overfitting 방지, 일반화 성능 향상
 - 3. hyperparameter α 를 사용하여 regularization 강도를 조절
- 차이점

Lasso	Ridge
L1 regularization	L2 regularization
계수 값을 0으로 줄일 수 있음.	계수 값을 작게 할 순 있으나, 0으로 만들지는 못함
변수 선택 가능	변수 선택 불가능
변수 간 상관관계가 낮은 상황에서 유용	변수 간 상관관계가 높은 상황에서 유용
특정 변수들이 예측에 중요한 역할을 하는 경우 유용	특정 변수들이 예측에 미치는 영향이 크게 다르지 않은 경우 유용

L1 regularization (Lasso)

- 계수의 절댓값 합을 최소화
- 최적화 과정에서 일부 계수가 0으로 수렴하는 경향이 있음. 이는 모델에서 해당 변수를 제외하는 역할을 함.
=> 변수 선택 과정이 포함됨

Find w, b such that minimizes ...

$$RSS_{Lasso}(w, b) = \sum_{i=1}^n (y_i - (w_i x_i + b))^2 + \alpha \sum_{j=1}^p |w_j|$$

L1 regularization (Lasso)

- 최적화 과정에서 일부 계수가 0으로 수렴하는 경향이 있음. 이는 모델에서 해당 변수를 제외하는 역할을 함.
=> 변수 선택 과정이 포함됨

$$w = (2xy - \alpha) / (2x^2)$$

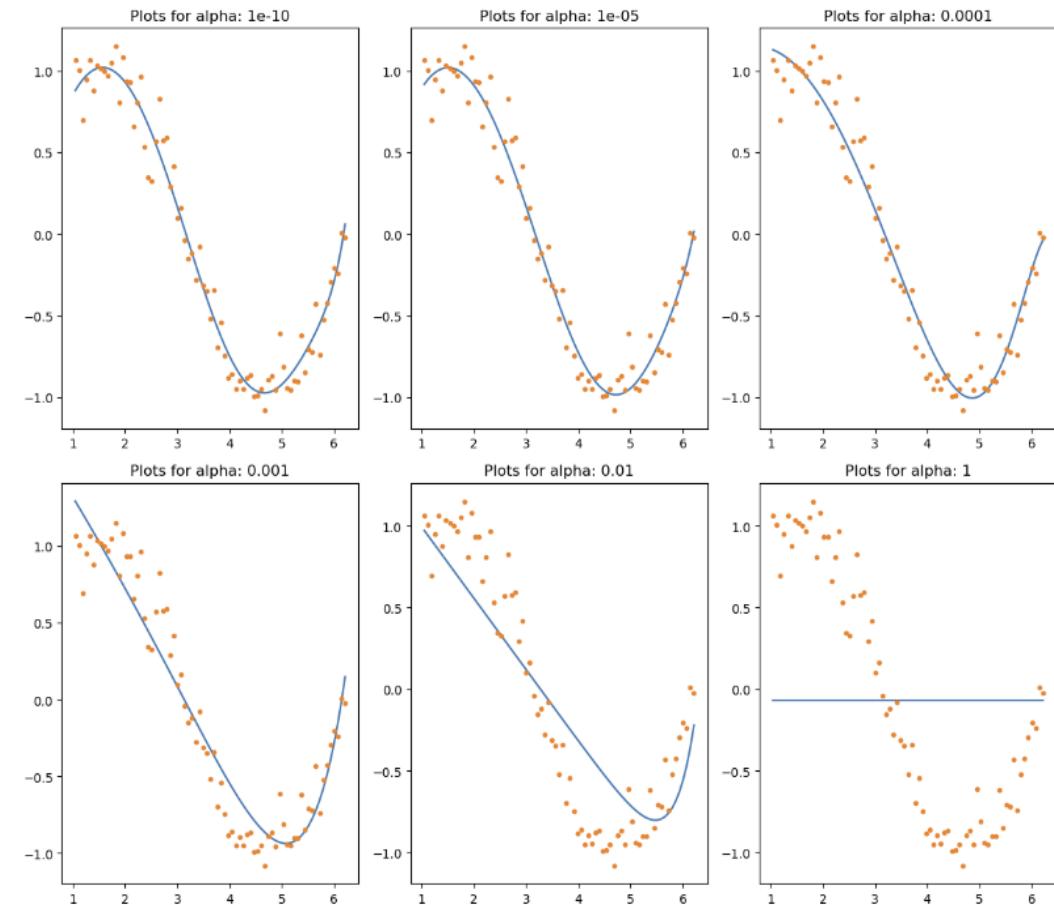
- 분자를 보면 α 를 빼는 부분이 있는데, 이는 분자를 0으로 만들 수 있음. 즉, w 가 0이 될 수 있음.
- 유용한 상황
 - 변수 간 상관관계가 낮을 때
 - 특정 변수들이 예측에 중요한 역할을 할 때

L1 regularization (Lasso)

- α = regularization 강도를 조절하는 hyperparameter
- α 가 커질수록 regularization 강도가 강해져서 계수의 크기를 작게 만듦
- α 가 0이면 regularization 없는 일반적인 선형 회귀와 동일

Find w, b such that minimizes ...

$$RSS_{Lasso}(w, b) = \sum_{i=1}^n (y_i - (w_i x_i + b))^2 + \alpha \sum_{j=1}^p |w_j|$$



L2 regularization (Ridge)

- 계수의 제곱 합을 최소화
- 계수가 커질수록 패널티가 기하급수적으로 증가
- 계수가 1보다 작을 때는 Lasso에 비해 규제를 덜 가함

Find w, b such that minimizes ...

$$RSS_{ridge}(w, b) = \sum_{i=1}^n (y_i - (w_i x_i + b))^2 + \alpha \sum_{j=1}^p w_j^2$$

L2 regularization (Ridge)

$$w = x y / (x^2 + \alpha)$$

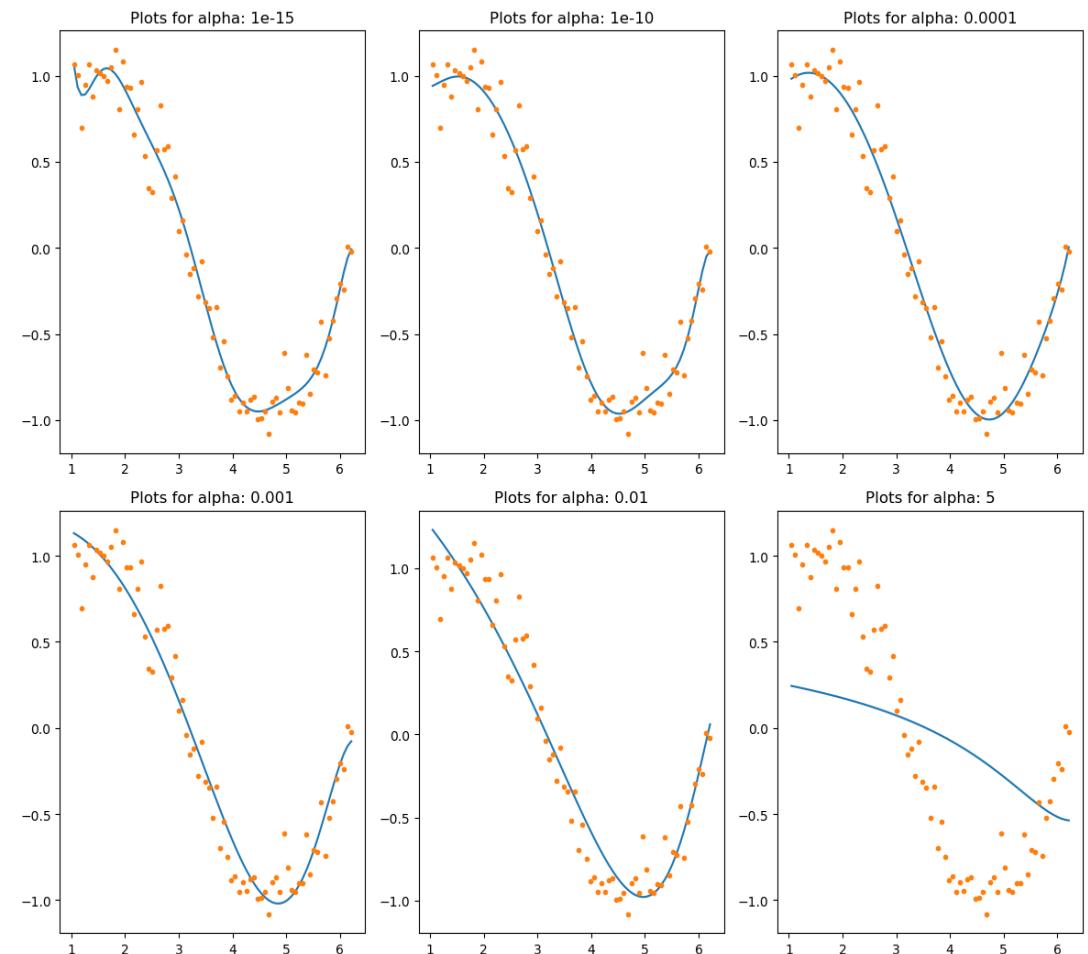
- 분모를 보면 α 를 더하는 부분이 있는데, 이는 $\alpha \rightarrow \infty$ 일 때만 분모를 0으로 만들 수 있음. 즉, w 를 작게 만들 순 있으나, 0으로 만들지는 못함.
- 유용한 상황
 - 변수 간 상관관계가 높을 때
 - 특정 변수들이 예측에 미치는 영향이 크게 다르지 않을 때

L2 regularization (Ridge)

- α = regularization 강도를 조절하는 hyperparameter
- α 가 커질수록 regularization 강도가 강해져서 계수의 크기를
작게 만듦
- α 가 0이면 regularization 없는 일반적인 선형 회귀와 동일

Find w, b such that minimizes ...

$$RSS_{ridge}(w, b) = \sum_{i=1}^n (y_i - (w_i x_i + b))^2 + \alpha \sum_{j=1}^p w_j^2$$



Elastic Net

- Lasso + Ridge
- 상관관계가 큰 변수들을 모두 선택하거나 제거 가능 -> 다수의 변수 간에 상관관계가 존재할 때 유용
- 두 개의 hyperparameter를 가짐 (Lasso에 대한 α_1 , Ridge에 대한 α_2)

$$ElasticNet = \sum_{i=1}^n (y_i - y(x_i))^2 + \alpha_1 \sum_{j=1}^p |w_j| + \alpha_2 \sum_{j=1}^p (w_j)^2$$

세션 끝!