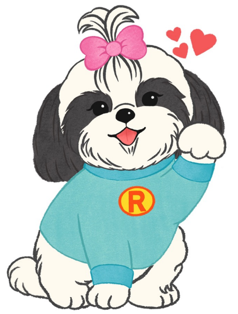
확률과 통계

섹션 - 5

강사 : James 쌤



통계 예측모형

키포인트

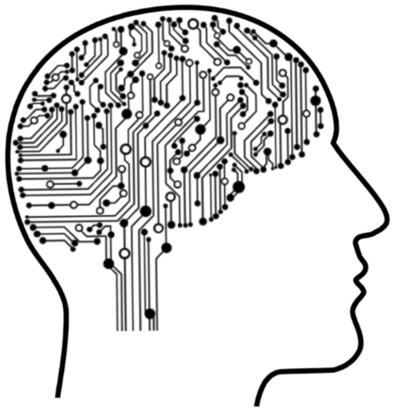
- 통계 예측모형 (머신러닝).
- 머신러닝의 유형.

머신러닝이란?

- 머신러닝이란 무었인가?

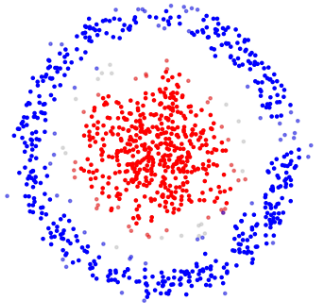
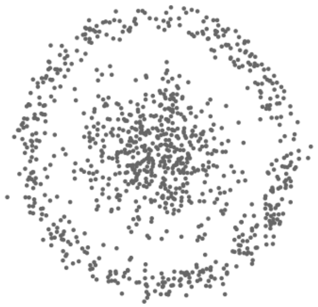
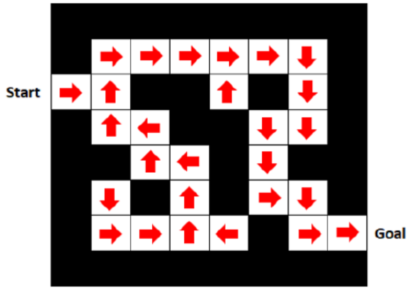
1) 데이터를 통해서 학습할 수 있는 알고리즘 또는 통계 모형.

2) 하드 코딩되지 않은 패턴을 데이터를 통해서 학습한다.



머신러닝의 유형

머신러닝의 유형

지도학습	비지도학습	강화학습
		
정답이 주어짐.	정답이 없음.	정책의 최적화.

머신러닝의 유형

- 학습용 데이터:

Y	X_1	X_2	$\cdots$	X_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

지도학습

X_1	X_2	$\cdots$	X_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

비지도학습

머신러닝의 유형

- 학습용 데이터:

종속변수	독립변수			
Y	X_1	X_2	$\cdots$	X_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

지도학습

독립변수			
X_1	X_2	$\cdots$	X_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

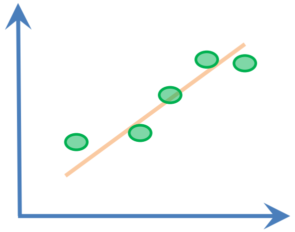
비지도학습

지도학습 머신러닝

지도학습의 세분화

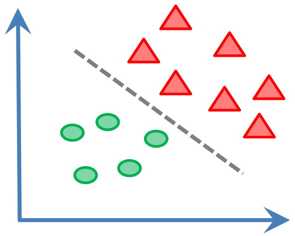
수치형 Y

$Y = 13.45,\ \ 73,\ \ 9.5,\ \ .....$



명목형 Y

$Y =$ *red*, *green*, *blue*, ...



지도학습 머신러닝

지도학습 머신러닝 세분화	Algorithm
Regression (회귀형) 수치형 Y	Linear Regression "선형회귀"
	K Nearest Neighbor Regressor
	Random Forest Regressor
	XGBoost Regressor
Classification (분류형) 명목형 Y	Logistic Regression "로지스틱회귀"
	K Nearest Neighbor Classifier
	Random Forest Classifier
	XGBoost Classifier

지도학습 머신러닝

지도학습 머신러닝 세분화	Algorithm
Regression (회귀형) 수치형 Y	Linear Regression "선형회귀"
	K Nearest Neighbor Regressor
	Random Forest Regressor
	XGBoost Regressor
Classification (분류형) 명목형 Y	Logistic Regression "로지스틱회귀"
	K Nearest Neighbor Classifier
	Random Forest Classifier
	XGBoost Classifier

자세히
배워본다!

지도학습 머신러닝

지도학습의 평가

수치형 Y

MSE, MAE, RMSE,

Correlation, 등.

명목형 Y

Confusion Matrix, Accuracy,

Precision, Recall, Specificity, 등.

선형회귀 원리 - Part 1

키포인트

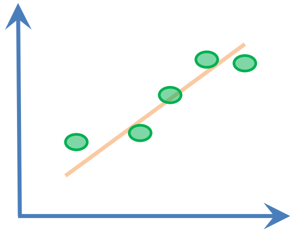
- 선형회귀의 원리.
- 선형회귀 모형의 해석.

지도학습 머신러닝

지도학습의 세분화

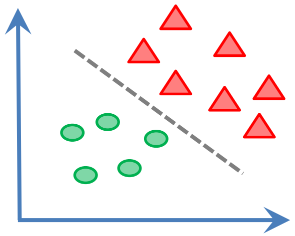
수치형 Y

$Y = 13.45,\ \ 73,\ \ 9.5,\ \ .....$



명목형 Y

$Y =$ red, green, blue, ...



지도학습 머신러닝

지도학습의 평가

수치형 Y

MSE, MAE, RMSE,

Correlation, 등.

명목형 Y

Confusion Matrix, Accuracy,

Precision, Recall, Specificity, 등.

선형회귀 개요

- 선형회귀는 대표적인 수치 예측 방법이다.
- 한 개 이상의 독립변수 (설명변수)가 있다: $X_1, X_2, \ldots, X_K$
- 한개의 종속변수 (반응변수)를 전제한다: Y "단변량"
- 선형 관계를 전제한다: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$
- 이외에도 여러 가지의 전제조건이 있다. ⇒ 잔차 분석 (later).

선형회귀 목적

1. 종속변수를 설명하는 독립변수를 밝혀낸다.

   예). 아파트의 가격은 면적, 위치, 방의 수 등으로 설명할 수 있다. (?)

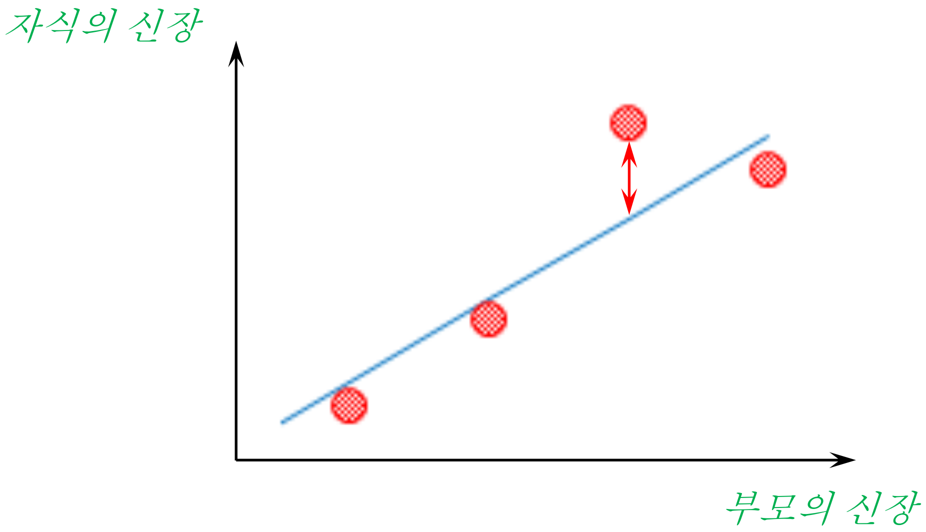
2. 독립변수 값이 데이터로 주어졌을 때 종속변수의 값을 예측한다.

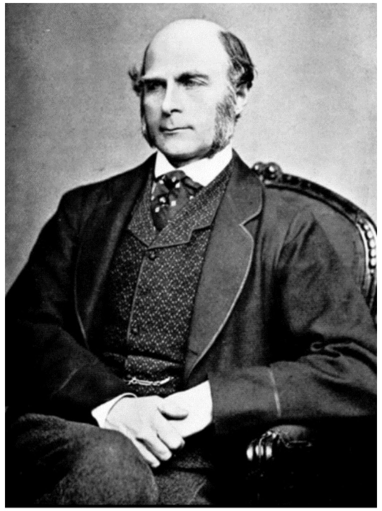
   예). 학습된 모형으로 아파트의 적정 가격을 알아 맞춘다.

역사적 배경

- 19세기 영국의 생물학자 Francis Galton이 “평균으로 돌아간다”의 의미로 “회귀”라는 용어를 처음 사용했다.

$\Rightarrow$ 신장에 있어서 부모와 자식 사이의 유전적 관계를 연구했다.





선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

종속변수

(데이터로 값이 주어짐/예측의 대상)

선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

독립변수

(데이터로 값이 주어짐)

선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

회귀 계수.

(학습을 통해서 계산됨)

선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

데이터 속의 패턴을 담는다.

선형회귀 원리

- 회귀모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

MPG

선형회귀 원리

- 회귀모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

MPG

N# of

Cylinders

선형회귀 원리

- 회귀모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Y: MPG

X_1: N# of Cylinders

X_2: HP

선형회귀 원리

- 회귀모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

MPG

N# of Cylinders

HP

Weight

선형회귀 원리

- 회귀모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Y	X_1	X_2	X_3	X_4
MPG	N# of Cylinders	HP	Weight	Auto or Manual

선형회귀 원리

- 오차변수 (white noise):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

평균[ε] = 0

표준편차[ε] = σ_ε

선형회귀 원리

- 공선성을 피해야 한다:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

$$Cor(X_i, X_j) \approx 0$$

for $i \neq j$

선형회귀 원리

- 공선성을 피해야 한다:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

공선성은 계수의 “분산 인플레”
문제를 일으킬 수 있다.

선형회귀 원리

- 회귀계수의 해석:

$$\Delta Y = \beta_1 \Delta X_1 + \cdots + \beta_i \Delta X_i + \cdots + \beta_K \Delta X_K$$

X변수가 ΔX 만큼 변동한다면

Y변수는 ΔY 만큼 반응한다.

선형회귀 원리

- 회귀계수의 해석:

$$\Delta Y = \beta_1 \Delta X_1 + \cdots + \beta_i \Delta X_i + \cdots + \beta_K \Delta X_K$$

1

β_i는 다른 X 변수들은 그대로 있으면서

X_i만 +1 증가할 때의 ΔY와 같다.

선형회귀 원리

- 회귀계수의 해석:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

절편 (intercept)

선형회귀 원리

- 회귀계수의 해석:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

절편 (intercept)

$X_1 = 0, X_2 = 0, \ldots$

선형회귀 원리

- 회귀계수의 해석:

$$Y = \beta_0 \underbrace{+ \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon}_{=0}$$

절편 (intercept)

선형회귀 원리

- 회귀계수의 해석:

$$Y = \beta_0$$

$\uparrow$

절편 (intercept)

“바닥”의 의미.

선형회귀 원리

- 회귀모형의 예 #2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

급여

선형회귀 원리

- 회귀모형의 예 #2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

급여

X_1 = 교육

X_2 = 경력

선형회귀 원리

- 회귀모형의 예 #2:

$\beta_1 > 0$

$\beta_2 > 0$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

급여

$X_1 =$ 교육

$X_2 =$ 경력

선형회귀 원리

- 회귀모형의 예 #2:

$\beta_1 > 0$

$\beta_2 > 0$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$X_1 =$ 교육

$X_2 =$ 경력

급여

$\beta_0 =$ 베이스 급여

선형회귀 원리 – Part 2

키포인트

- 선형회귀 "최소자승법" OLS 해.
- 선형회귀 학습과 예측.
- 가변수 또는 더미변수 (dummy variable).

선형회귀 원리

- 회귀모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$y_j = \beta_0 + \beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_K x_{j,K} + \varepsilon_j$$

$j \in [1, n]$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$\overrightarrow{Y} = \tilde{X}\,\overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$\vec{Y} = \tilde{X}\,\vec{\beta} + \vec{\varepsilon}$$

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$\overrightarrow{Y} = \tilde{X}\,\overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

$$\tilde{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,K} \\ 1 & x_{2,1} & \cdots & x_{2,K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,K} \end{pmatrix}$$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$\overrightarrow{Y} = \tilde{X}\,\overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

$$\overrightarrow{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

선형회귀 해

- 벡터와 행렬 사용 표기:

$$\overrightarrow{Y} = \tilde{X}\,\overrightarrow{\beta} + \overrightarrow{\varepsilon}$$

$$\overrightarrow{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

선형회귀 해

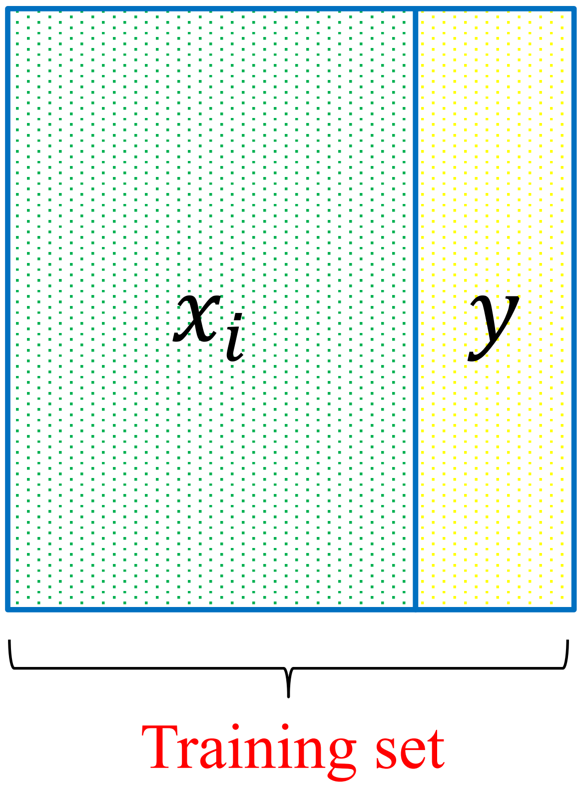
- 선형회귀의 OLS (**O**rdinary **L**east **S**quares) 해:

$$\vec{\beta} = \left[\left(\tilde{X}^t \tilde{X}\right)^{-1} \tilde{X}^t\right] \vec{Y}$$

$$\Uparrow$$

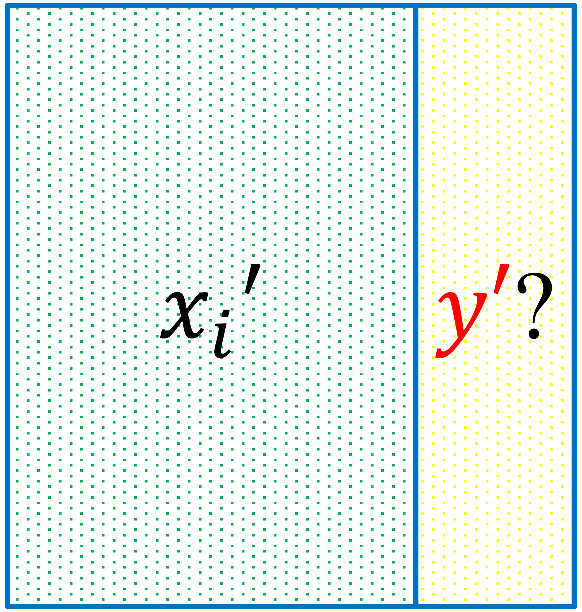
$\|\vec{\varepsilon}\|^2$를 최소화 하는 계수벡터이다.

선형회귀 학습



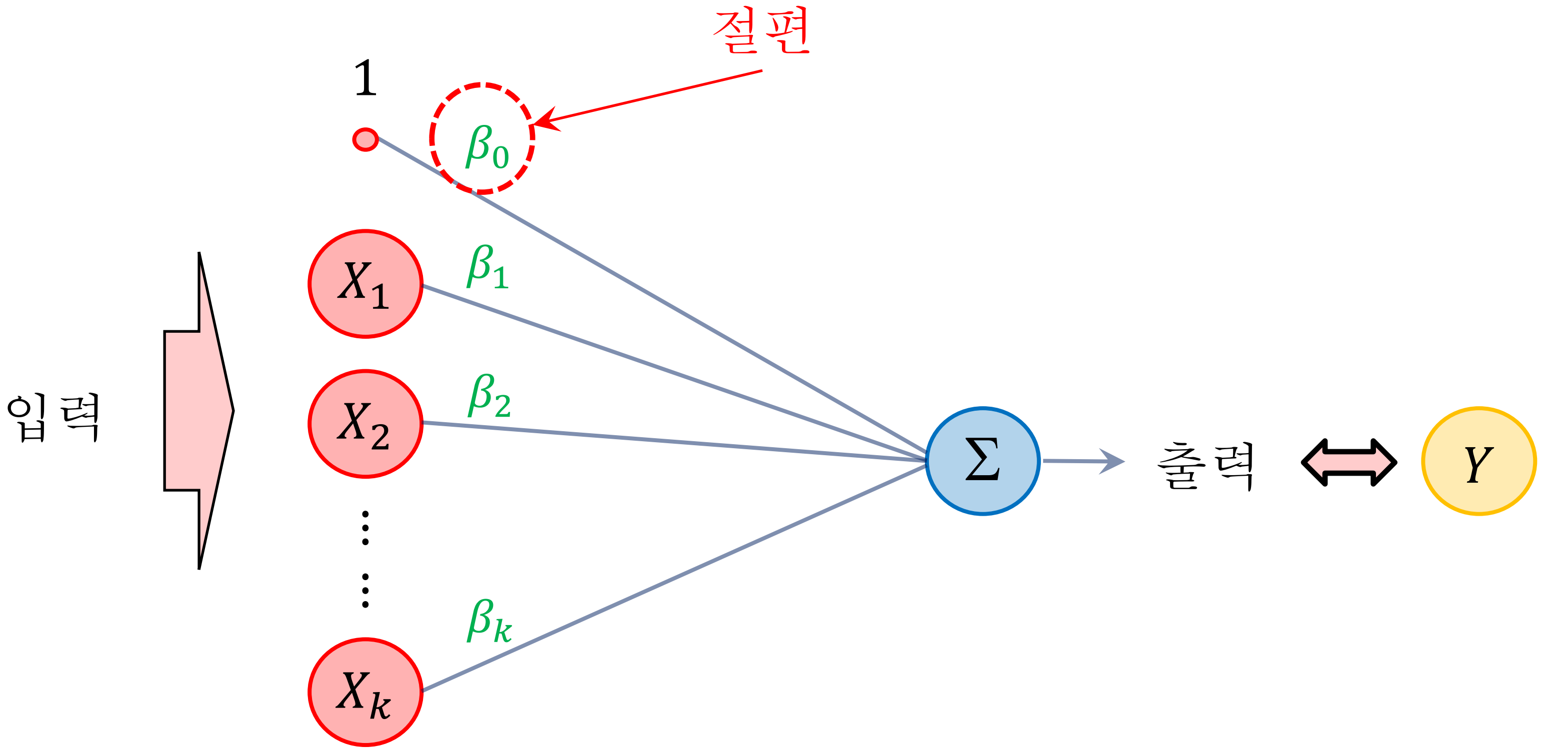
모형의 파라미터, 즉 $\{\beta_i\}$를 학습용 데이터를 사용하여 계산해 놓는다.

선형회귀 예측

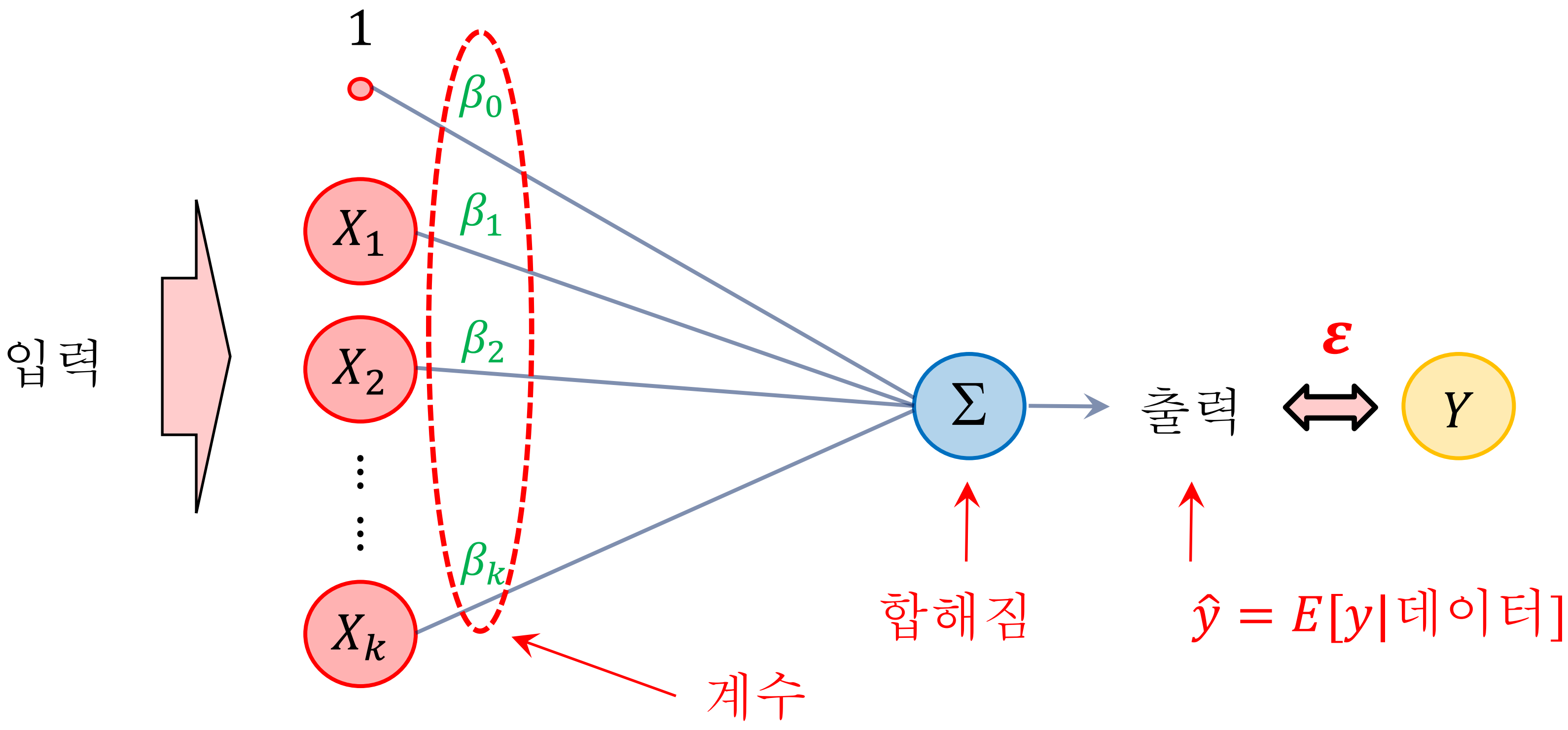


독립변수의 값이 새롭게 주어졌을 때 $\{x_i'\}$, 모르는 상태인 종속변수의 값 y' 을 계산을 통해서 알아낸다.

선형회귀 예측



선형회귀 예측



1
β_0
β_1
β_2
β_k
입력
X_1
X_2
X_k
Σ
출력
ε
Y
계수
합해짐
$\hat{y} = E[y|\text{데이터}]$

선형회귀 예측

- 독립변수의 값 $x_1', x_2', \ldots, x_k'$가 데이터로 주어졌을 때 다음 수식을 사용해서 종속변수의 예측값 $\hat{y} = E[y|\text{데이터}]$를 구할 수 있다.

$$\hat{y} = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \cdots + \beta_K x_K'$$

선형회귀 예측

- 독립변수가 단 하나 뿐인 특수한 경우에는 예측값의 95% 신뢰구간을 계산할 수 있다.

$$\left[\hat{y} - \boldsymbol{qt}(0.975, n-2)\,\sigma_{\hat{y}}\,,\hat{y} + \boldsymbol{qt}(0.975, n-2)\,\sigma_{\hat{y}}\right]$$

$$\sigma_{\hat{y}} = RMSE \times \sqrt{\left(\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum(x - \bar{x})^2}\right)}$$

$$\Uparrow$$

$qt()$은 R의 분위수 함수, RMSE = Root Mean Square Error.

가변수/더미변수

- 가변수 또는 더미변수는 0과 1만을 값으로 갖는 변수이다. ⇒ switch on/off 의 역할.
- 명목형 변수를 모형에 추가하면 유형의 가지수 – 1 개의 더미변수 생성됨.

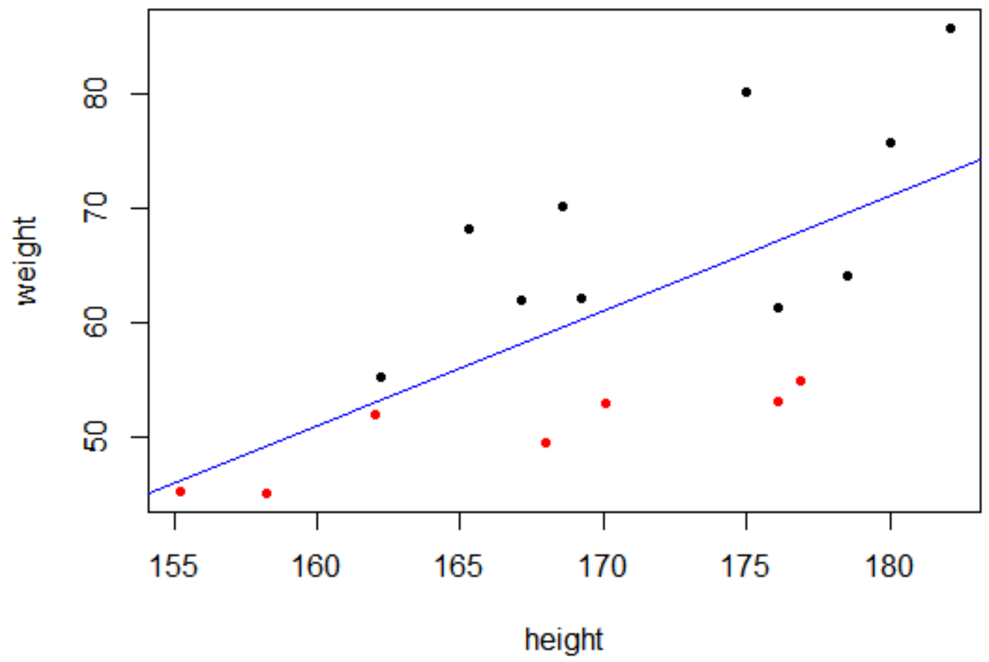
예). "남자", "여자"와 같이 두 개의 유형을 값으로 갖는 "gender" 변수는 "gender_여자"라는 한 개의 더미변수를 생성한다.

예). "setosa", "versicolor", "virginica"와 같이 세 개의 유형을 값으로 갖는 "Species" 변수는 "Species_versicolor", "Species_virginica" 라는 두 개의 더미변수를 생성한다.

가변수/더미변수

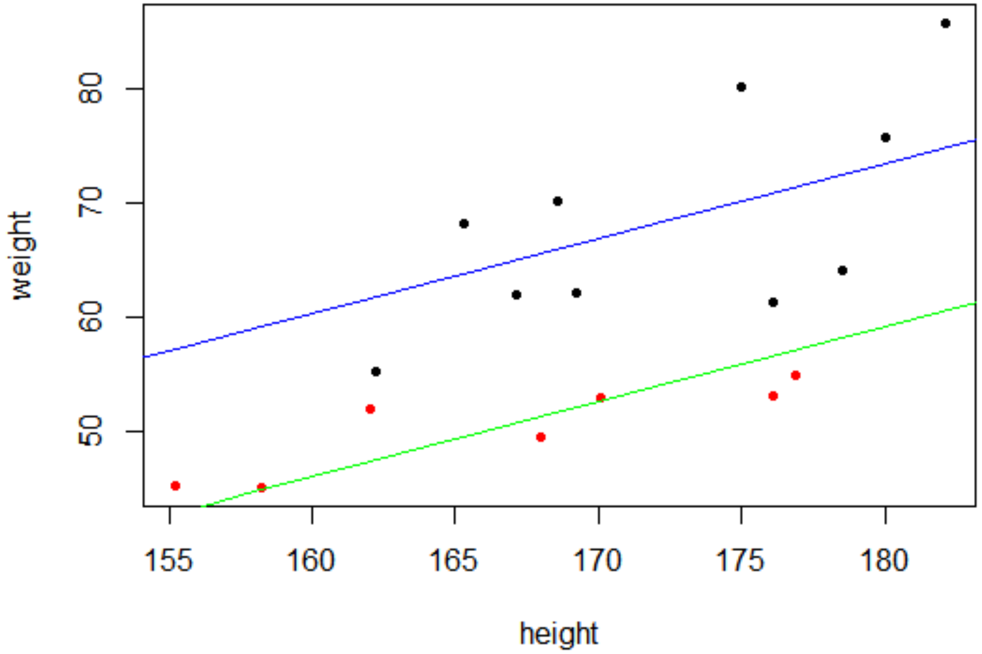
- 더미변수가 (다른 변수와 상호작용하지 않고) 독립적으로 추가되면 해당 유형의 절편을 올리거나 내려주는 역할을 한다.
- 상호작용하는 더미변수는 해당 유형의 기울기를 조절해 주는 역할을 한다.

가변수/더미변수



더미변수 없음: weight ~ height

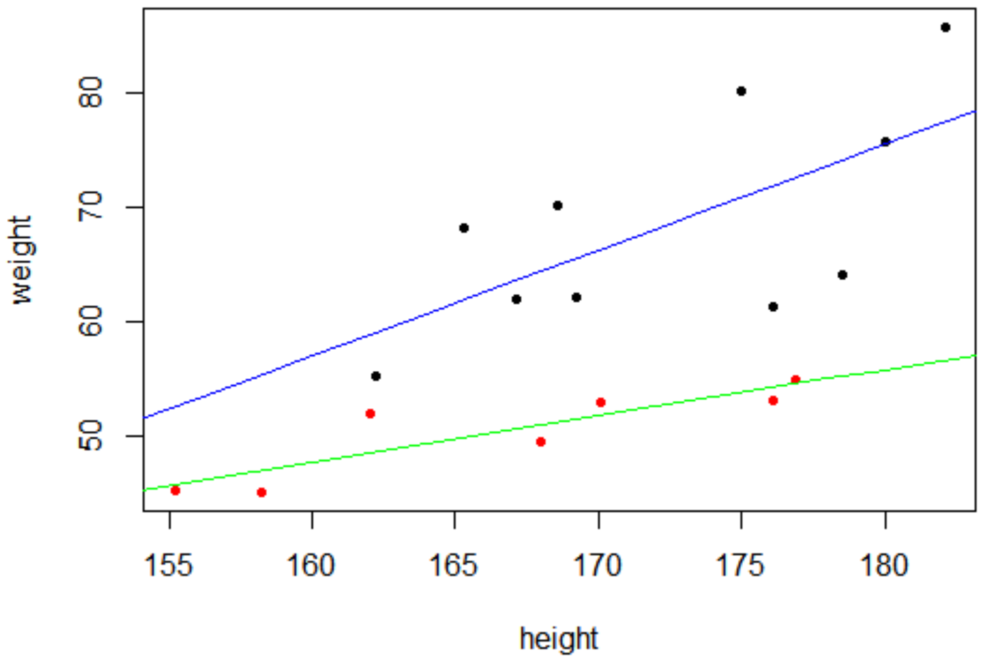
가변수/더미변수



더미변수 있음: weight ~ height + gender

R^2증가, MSE 감소

가변수/더미변수



상호작용하는 더미변수 있음: weight ~ height * gender

R^2증가, MSE 감소

회귀모형의 진단과 선별

키포인트

- 선형회귀 모형의 진단.
- t 검정을 적용한 회귀 계수의 유의성 확인.
- F 검정을 적용한 회귀 모형의 설명력 확인.
- 결정계수 R^2, MSE, VIF 등과 같은 진단 수치 확인.
- 정보량을 활용한 선형회귀 모형의 선별.

선형회귀 진단: t 검정

Question : 모형의 설명변수들은 통계적으로 의미 있나?

⇒ 개개의 회귀 계수에 대한 t 검정.

선형회귀 진단: t 검정

- 개개의 회귀 계수에 대한 양측검정 (t 검정)을 실행한다.

  귀무가설 $\boldsymbol{H_0} : \beta_i = 0$

  대립가설 $\boldsymbol{H_1} : \beta_i \neq 0$

  $\Rightarrow$ t 검정 통계량과 p-값 사용.

선형회귀 진단: t 검정

- 개개의 회귀 계수에 대한 양측검정 (t 검정)을 실행한다.

  귀무가설 $\boldsymbol{H_0} : \beta_i = 0$

  대립가설 $\boldsymbol{H_1} : \beta_i \neq 0$

  $\Rightarrow$ t 검정 통계량 $= \dfrac{\widehat{\beta_i}}{\beta_i\text{의 표준오차}}$

  $\Rightarrow$ p-값이 임계치 이하인 경우 ($<$ 0.05), $\boldsymbol{H_0}$ 기각 후 $\boldsymbol{H_1}$채택.

  X_i를 모형에 포함시키는 것이 **정당화** 된다.

선형회귀 진단: F 검정

Question : 회귀 모형은 설명력을 제공하나?

선형회귀 진단: F 검정

Question : 회귀모형의 독립변수 중 최소 한 개라도 종속변수를 설명하는 역할을 하고 있나?

선형회귀 진단: F 검정

- 모든 회귀 계수에 대한 F 검정을 실행한다.

  귀무가설 $\boldsymbol{H_0}$: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$.

  대립가설 $\boldsymbol{H_1}$: 적어도 한 개 이상의 β_i가 0과 다르다.

  $\Rightarrow$ F 검정 통계량과 p-값 사용.

선형회귀 진단: F 검정

- 모든 회귀 계수에 대한 F 검정을 실행한다.

  귀무가설 $\boldsymbol{H_0}$: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$.

  대립가설 $\boldsymbol{H_1}$: 적어도 한 개 이상의 β_i가 0과 다르다.

  $\Rightarrow$ F 검정 통계량 $= \dfrac{\text{설명할 수 있는 오류 (분산)}}{\text{설명할 수 없는 오류 (분산)}}$

  $\Rightarrow$ p-값이 임계치 이하인 경우 (< 0.05), $\boldsymbol{H_0}$ 기각 후 $\boldsymbol{H_1}$채택.

  회귀 모형은 최소 이상의 설명력 있다.

선형회귀 진단: 결정계수

- 결정계수 R^2 는 대표적인 진단 척도중의 하나이다.
- $0 < R^2 < 1$이며 R^2이 1에 가까울 수록 좋다.

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

with $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

선형회귀 진단: 결정계수

- 모형이 복잡해 질수록 일단 R^2은 증가한다.

  ⇒ R^2만을 기준으로 모형을 만들면 과적합 현상이 쉽게 발생하니 주의한다.

  ⇒ "조정 결정계수" (adjusted R^2)가 선호된다.

- 독립변수가 **하나 뿐인** 경우에는 R^2 는 X와 Y 사이의 상관계수의 제곱과 값이 같다.

$$R^2 = Cor(X,Y)^2$$

선형회귀 진단: MSE, RMSE, MAE

- MSE와 RMSE는 예측값과 실제값 사이의 차이를 나타낸다. ⇒ 작을수록 좋다.

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

$$RMSE = \sqrt{MSE}$$

- MAE도 MSE와 유사한 의미를 갖는다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

선형회귀 진단: VIF

- 다중공선성의 정도는 개개 설명변수의 VIF (Variance Inflation Factor)를 사용하여 가능해 볼 수 있다.

  VIF > 5 : 강한 다중공선성 존재.

  VIF > 10 : 심각한 수준의 다중공선성 존재.

- VIF 수치가 큰 경우 “모형 간추리기”가 필요할 수 있다.

선형회귀 진단: VIF

- 개개 설명변수 X_i에 대한 VIF는 다음과 같은 방식으로 구한다.

1) 변수 X_i를 종속변수의 역할에 놓고 나머지 설명변수로 선형회귀식을 만든다:

$$X_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \cdots + \varepsilon$$

2) 해당 결정계수 R_i^2를 사용하여 VIF_i를 계산한다:

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

정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{n} + 2\frac{p}{n}$$

$$BIC = -2\frac{Log\ likelihood}{n} + p\frac{Ln(n)}{n}$$

$$Log\ likelihood = -\frac{n}{2}\left(1 + Ln(2\pi) + Ln\left(\frac{SSE}{n}\right)\right)$$

정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{n} + 2\frac{p}{n}$$

$$BIC = -2\frac{Log\ likelihood}{n} + p\frac{Ln(n)}{n}$$

$\Rightarrow$ AIC (또는 BIC)를 최소화 하려고 한다.

$\Rightarrow$ AIC (또는 BIC)는 두 개의 상반된 추세의 합이다.

정보량과 모형 선별

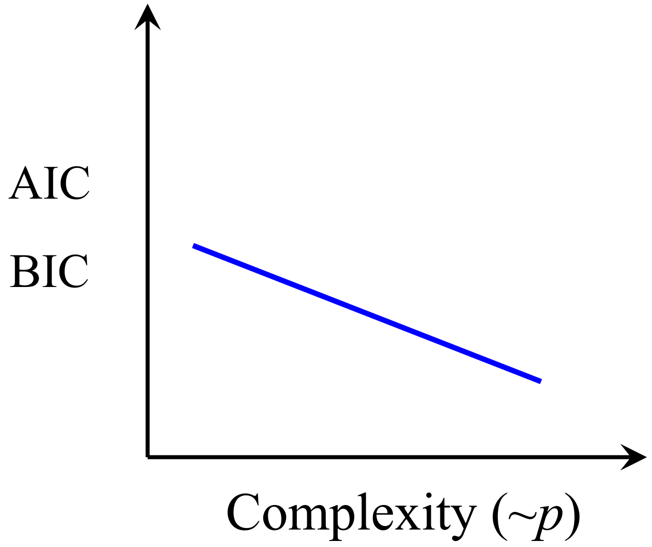
- 정보량 (Information Criteria):

$$AIC = \boxed{-2\frac{Log\ likelihood}{n}} + 2\frac{p}{n}$$

$$BIC = \boxed{-2\frac{Log\ likelihood}{n}} + p\frac{Ln(n)}{n}$$

~ *-Log likelihood*

모형이 복잡할 수록 감소.



정보량과 모형 선별

- 정보량 (Information Criteria):

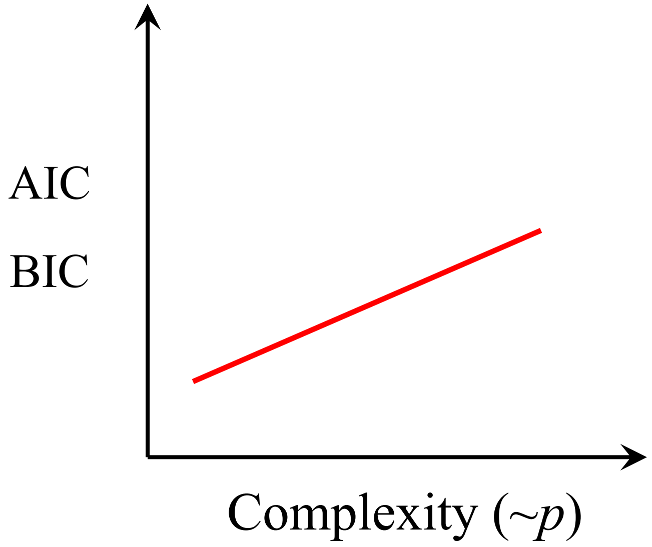
$$AIC = -2\frac{Log\ likelihood}{n} + \boxed{2\frac{p}{n}}$$

$$BIC = -2\frac{Log\ likelihood}{n} + \boxed{p\frac{Ln(n)}{n}}$$

$\sim p$

모형이 복잡할 수록 증가.

p는 모형 파라미터의 수.

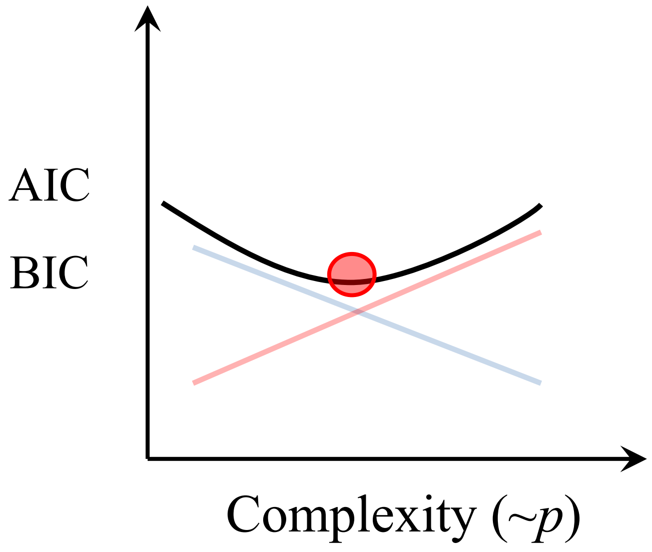


정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{n} + 2\frac{p}{n}$$

$$BIC = -2\frac{Log\ likelihood}{n} + p\frac{Ln(n)}{n}$$



$\Rightarrow$ 합이 최소인 **최적점**이 있다.

정보량과 모형 선별

- 회귀모형의 선별 방법:

  ⇒ R^2은 1에 가까워져야 한다.

  ⇒ AIC가 감소하는 방향으로 최적화를 진행한다.

  ⇒ 모형이 잘못된 방향으로 변경되면, AIC는 감소하는 대신에 증가한다.

Stop!

잔차와 레버리지 분석

키포인트

- 잔차와 레버리지를 사용한 영향력 분석.
- 쿡의 거리.
- 잔차 분석을 통한 선형회귀의 전제 조건 확인.

잔차와 레버리지 개요

- 잔차와 레버리지를 통한 영향력 분석을 하는 이유.

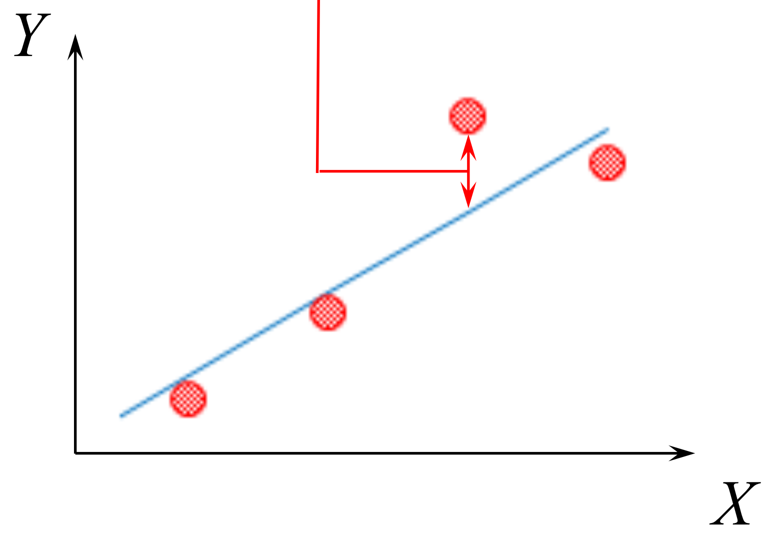
$\Rightarrow$ 잔차 : 종속변수 Y 에서 특이값을 발견할 수 있다.

$\Rightarrow$ 레버리지 : 설명변수 X 에서 특이값을 발견할 수 있다.

$\Rightarrow$ 가장 “영향력”이 큰 데이터 포인트를 찾을 수 있다.

잔차

- 잔차는 종속변수에 대한 예측값 $\hat{Y}$와 실제값 Y 사이의 차이이다.



⇒ 그러므로 Y 의 특이값을 쉽게 찾아낼 수 있다.

잔차

- 잔차분석을 통한 선형회귀 전제 조건을 확인할 수 있다:

  $\Rightarrow$ 선형성: 종속변수는 설명변수의 선형조합으로 설명이 가능하다.

  그러므로, 잔차에는 추세가 없어야 한다.

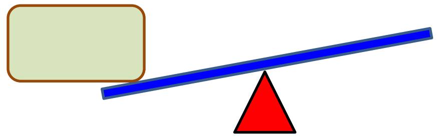
  $\Rightarrow$ 독립성: 잔차는 순서와 상관없이 독립적이어야 한다.

  $\Rightarrow$ 정상성: 잔차의 분포가 정규분포를 따라야 한다.

  $\Rightarrow$ 등분산성: 잔차의 분산이 순서와 무관하게 일정해야 한다.

레버리지

- 레버리지는 X값이 중앙에서 얼마나 멀리 떨어져 있는지를 나타낸다.
- 레버리지가 크다는 것은 회귀계수에 영향이 크다는 의미이다.

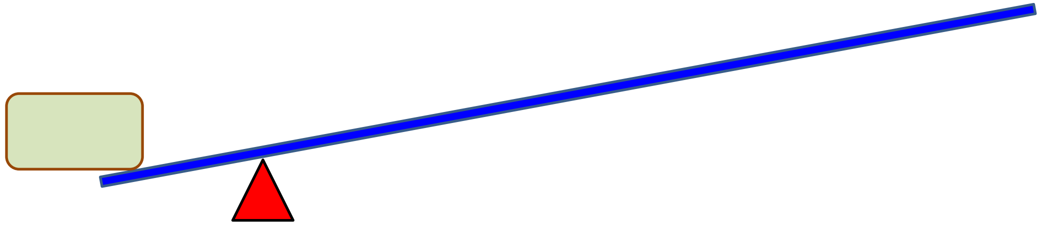


Short leverage

⇒ 그러므로 X의 특이값을 쉽게 찾아낼 수 있다.

레버리지

- 레버리지는 X값이 중앙에서 얼마나 멀리 떨어져 있는지를 나타낸다.
- 레버리지가 크다는 것은 회귀계수에 영향이 크다는 의미이다.



Long leverage

⇒ 그러므로 X의 특이값을 쉽게 찾아낼 수 있다.

레버리지

- i 번째 관측값의 레버리지:

$$Leverage = H_{ii} \longleftarrow \tilde{H} = \tilde{X}\left(\tilde{X}^t\tilde{X}\right)^{-1}\tilde{X}^t$$

- “Sum rule”:

$$\sum_{i=1}^{n} H_{ii} = p \longleftarrow \text{파라미터의 개수}$$

레버리지

- 레버리지의 상대적 크기를 판단하는 기준:

$$\text{평균의 레버리지} \cong \frac{p}{n}$$

$$\text{큰 레버리지} > \frac{p}{n}$$

$$\text{작은 레버리지} < \frac{p}{n}$$

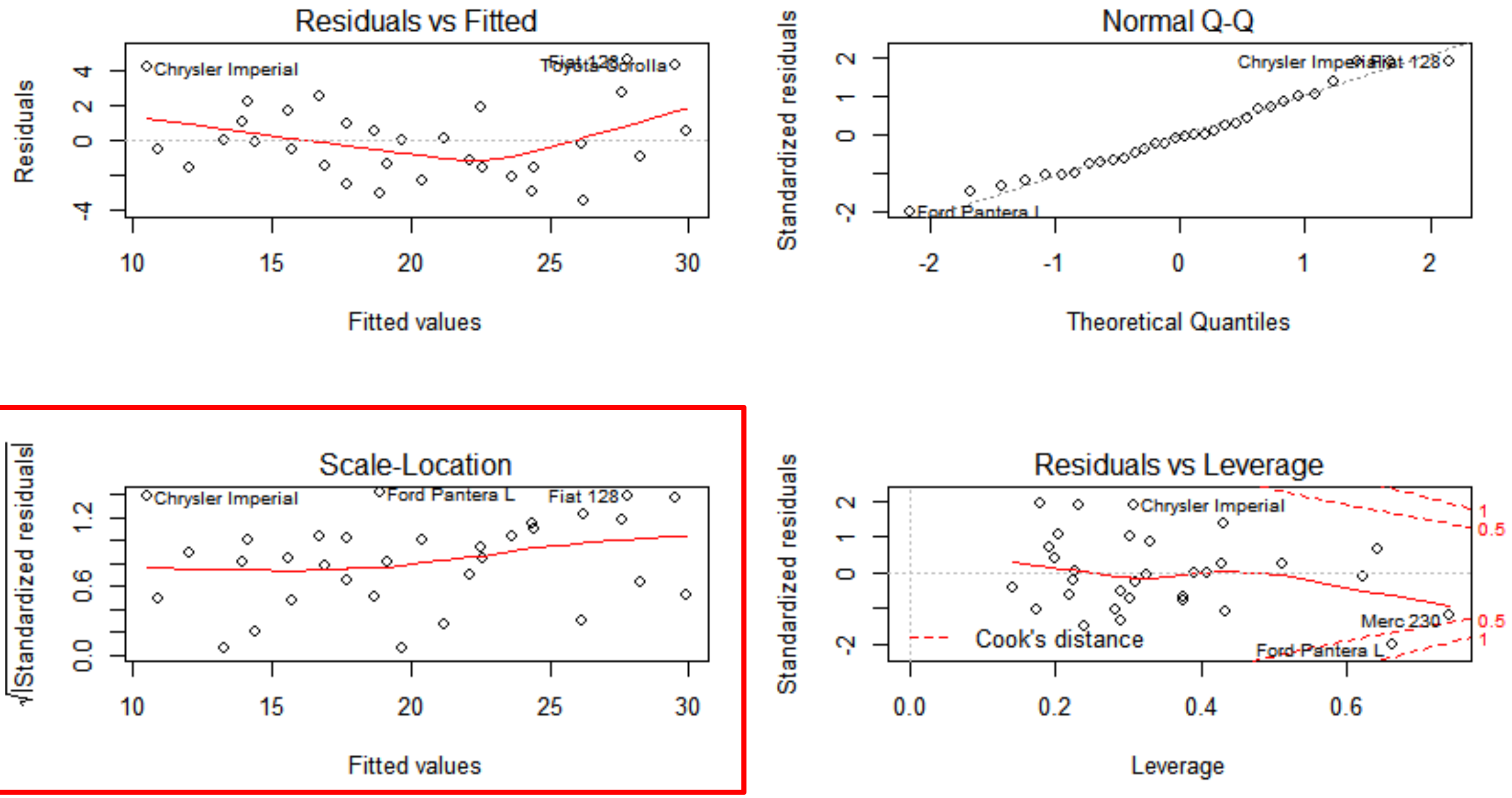
쿡의 거리

- i 번째 관측값의 쿡의 거리 (Cook's Distance):

$$D_i = \frac{{e_i}^2}{P \times MSE}\left[\frac{H_{ii}}{(1 - H_{ii})^2}\right]$$

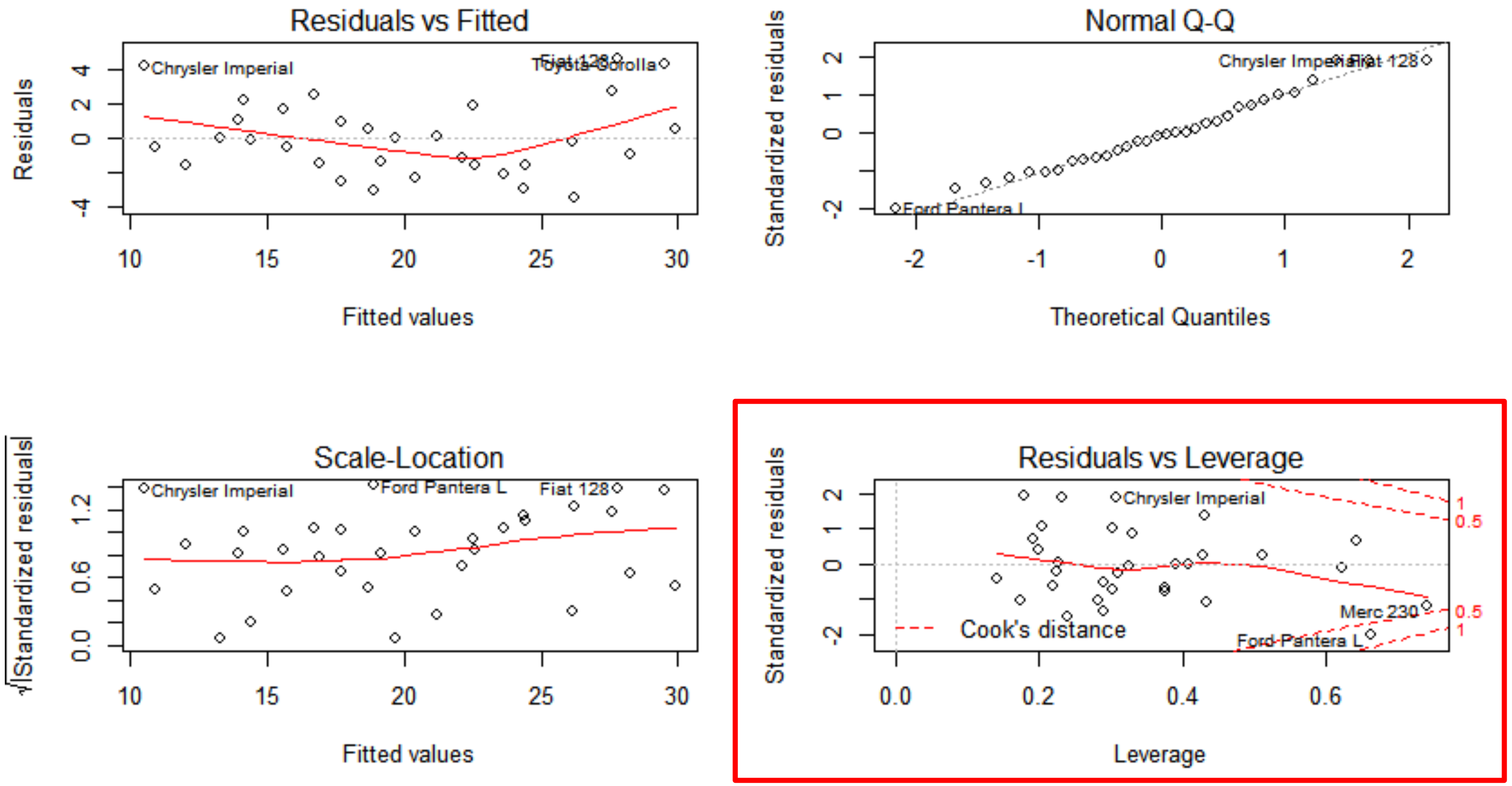
- 쿡의 거리는 전체적인 영향력을 나타내어 준다.
- 잔차와 레버리지의 "혼합된 개념"과도 유사하다.

시각화를 통한 외상치와 영향력 확인



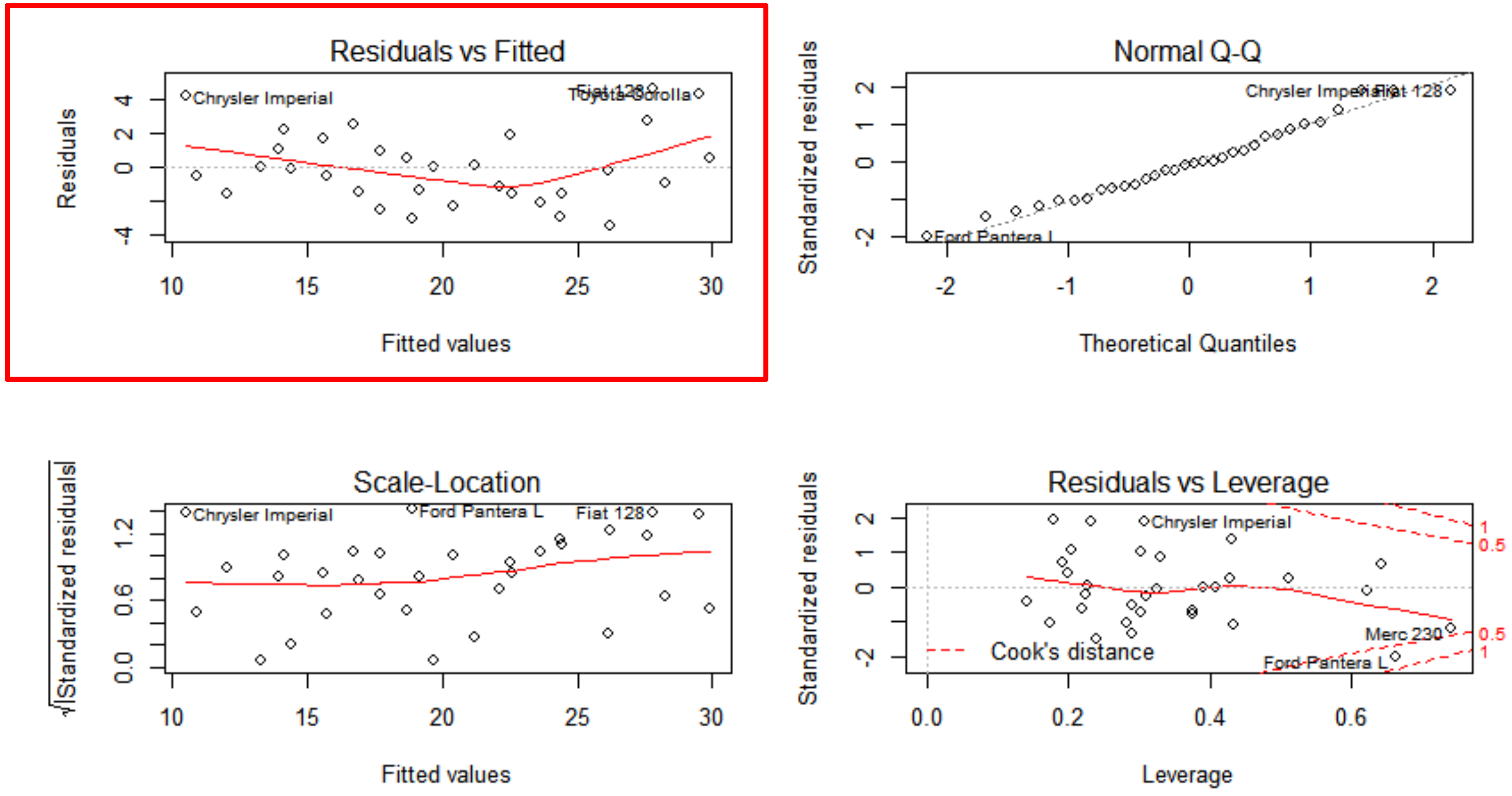
⇒ 시각적으로 외상치 (표준화된 잔차) 확인.

시각화를 통한 외상치와 영향력 확인



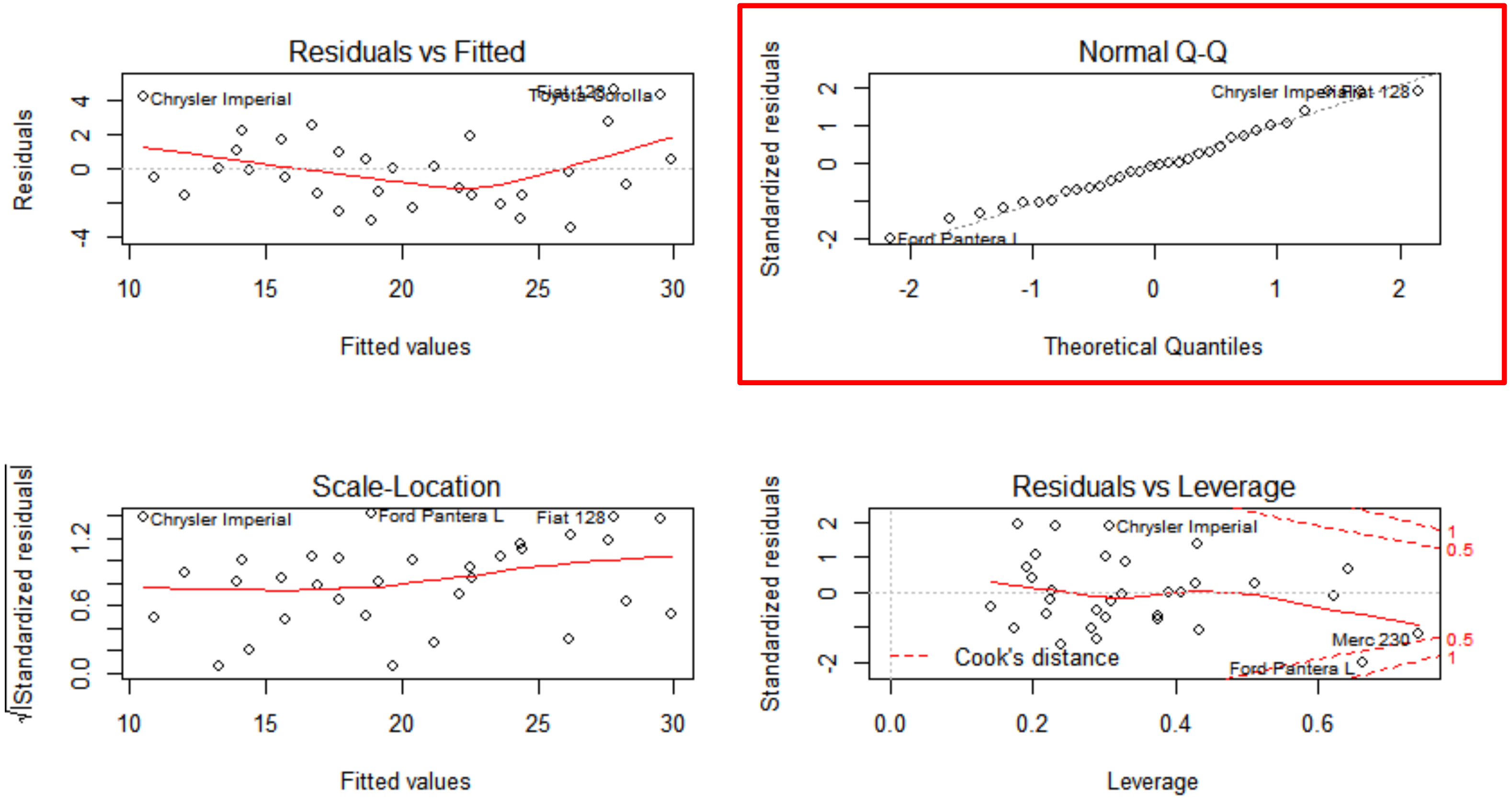
⇒ 시각적으로 영향력 (레버리지, 쿡의 거리) 확인.

시각화를 통한 외상치와 영향력 확인



⇒ 시각적으로 선형성, 독립성, 등분산성 확인.

시각화를 통한 외상치와 영향력 확인



⇒ QQ plot을 사용해서 시각적으로 정상성 확인.

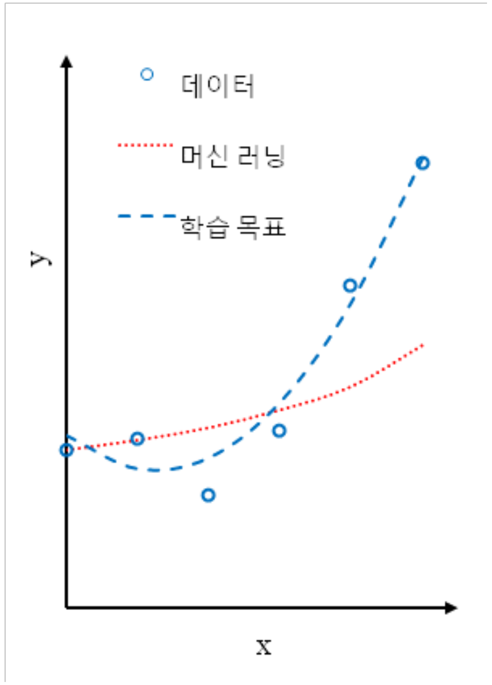
회귀분석의 유형

키포인트

- 편향 오류와 분산 오류.
- Ridge 회귀.
- Lasso 회귀.
- 다항식 회귀.
- 푸아송 회귀.

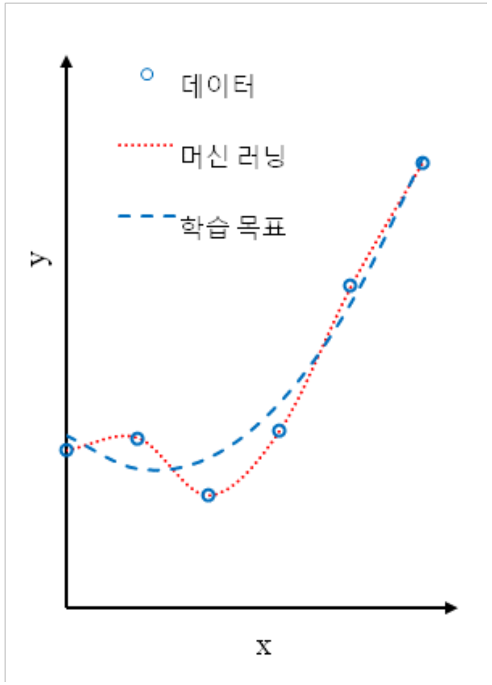
편향 오류

- 편향 오류 (bias error) 또는 과소적합 오류 (underfitting error).
- 모형이 편향적 즉 과하게 단순해서 발생하는 오류의 유형이다.



분산 오류

- 분산 오류 (variance error) 또는 과적합 오류 (overfitting error).
- 모형이 과하게 복잡해서 발생하는 오류이며 매개변수 최적화가 어려워 진다.



분산 오류

- 분산 오류 (variance error) 또는 과적합 오류 (overfitting error).
- 모형이 과하게 복잡해서 발생하는 오류이며 매개변수 최적화가 어려워 진다.
- In-sample 오류는 작지만 Out-of-sample 오류는 큰 경우이다.

⇒ In-sample 오류: 같은 데이터셋으로 트레이닝과 테스팅 진행.

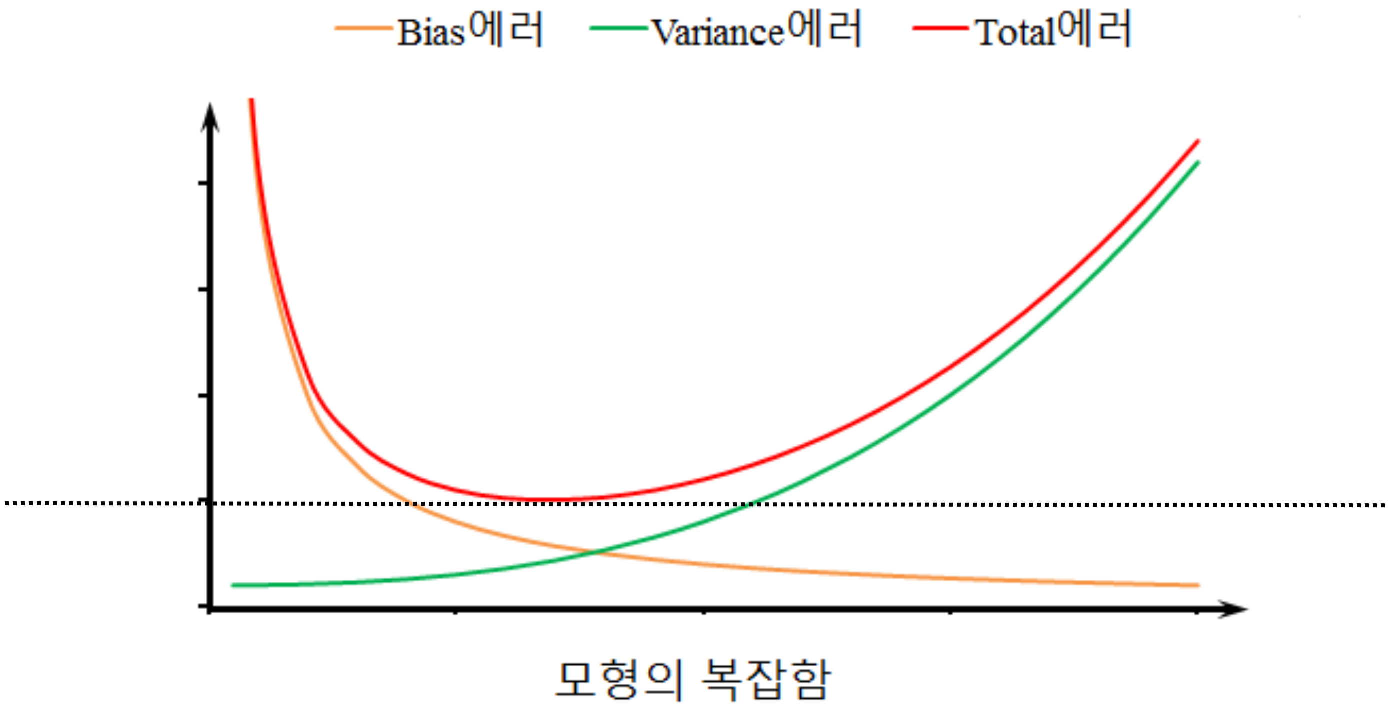
⇒ Out-of-sample 오류: 별도의 트레이닝 데이터셋과 테스팅 데이터셋.

토탈 오류

토탈 오류 = 편향 오류 + 분산 오류 + 상수

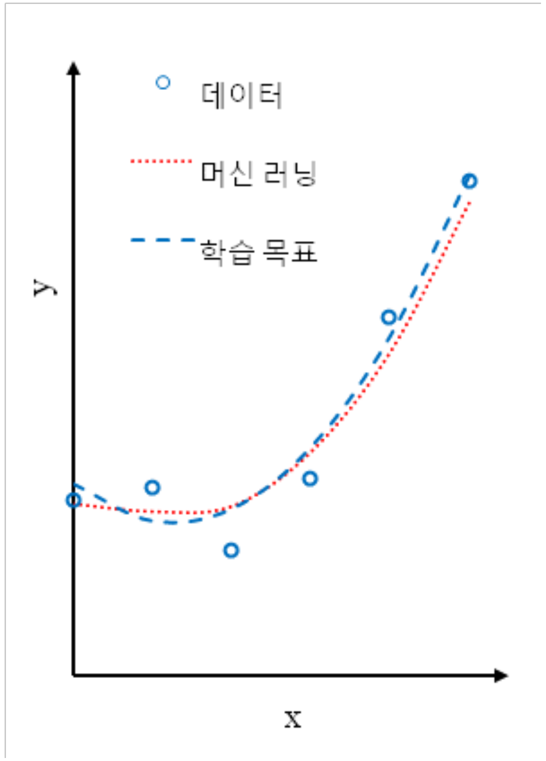
Out-of-Sample 시험 오류의 최소화

- 모형의 복잡함 (complexity)에는 최적점 (optimal point)이 있다.



Out-of-Sample 시험 오류의 최소화

- 다음은 최적화된 모형의 예시이다.



Ridge 회귀

- OLS해는 $\|\vec{\varepsilon}\|^2$를 최소화 하는 회귀계수 벡터를 구한다.
- Ridge회귀에서는 다음 손실함수를 최소화 한다. (L2 정규화)

$$L = \|\vec{\varepsilon}\|^2 + \boldsymbol{\lambda} \sum_{i=0}^{K} \beta_i^2$$

- λ는 양수로서 크면 클수록 분산오류를 줄이며 편향오류를 증가시킨다.
- 과적합 (overfitting)의 상황이 의심될 때 사용한다.
- 회귀계수의 절대값은 억제되지만 정확하게 0이 되지는 않는다.

Ridge 회귀

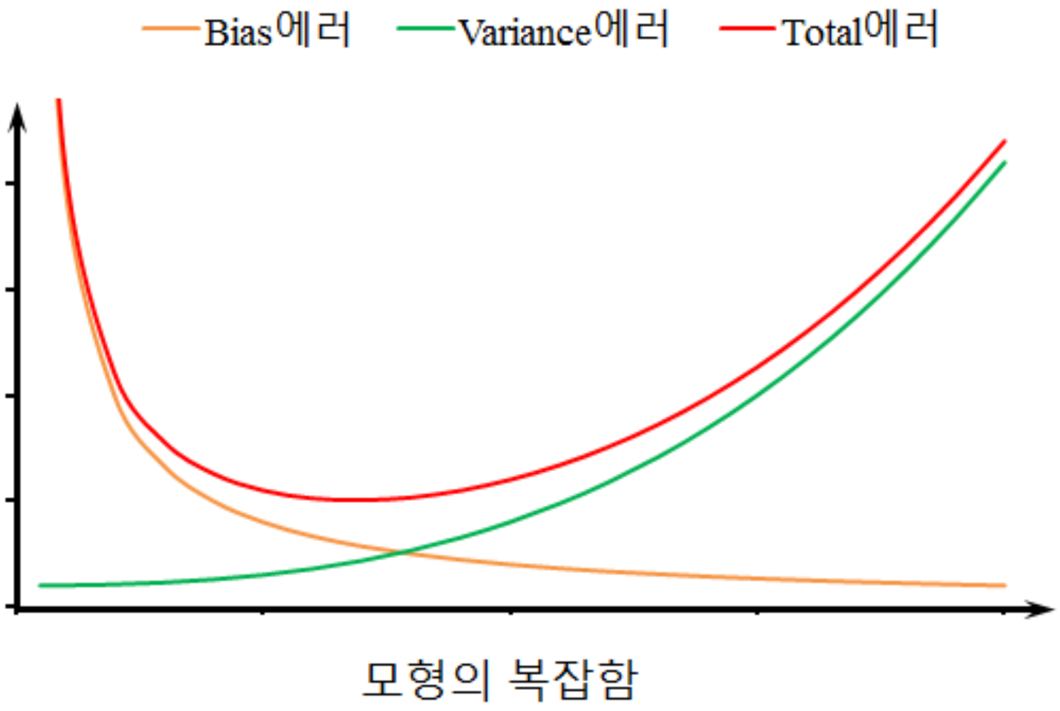
- λ는 양수이며 크면 클수록 회귀계수의 증가를 억제한다.

$$L = \|\vec{\varepsilon}\|^2 + \boldsymbol{\lambda} \sum_{i=0}^{K} \beta_i^2$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

Ridge 회귀

- 편향오류와 분산오류 사이의 trade-off 관계를 상기해 본다.



Lasso 회귀

- Lasso회귀에서는 다음 손실함수를 최소화 한다. (L1 정규화)

$$L = \|\vec{\varepsilon}\|^2 + \lambda \sum_{i=0}^{K} |\beta_i|$$

- Ridge회귀와 마찬가지로 과적합 (overfitting)의 상황이 의심될 때 사용.
- λ가 과하게 크면 편향오류의 증가가 분산오류의 감소를 상쇄하고도 남을 수 있으니 주의한다.
- 회귀계수가 정확하게 0이 될 수 있다.

다항식 회귀

- 다음과 같은 다항식을 사용하여 X와 Y 사이의 관계를 모형화 한다.

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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

- 주의할 점은 단 하나의 독립변수 X가 있다는 것이다.
- 다항식항은 I(X^2), I(X^3), 등과 같이 R 수식에 추가한다.

푸아송 회귀

- 종속변수 Y가 횟수 (count)를 나타내는 경우에 사용한다. 다음 관계를 전제한다.

$$Log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

- 푸아송 확률분포:

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$\Rightarrow$ 평균 $= \lambda$

$\Rightarrow$ 분산 $= \lambda$

$\Rightarrow$ 표준편차 $= \sqrt{\lambda}$

끝