Statistics II

for Machine Learning

Chap. 4: 로지스틱 회귀분석 & 다중공선성

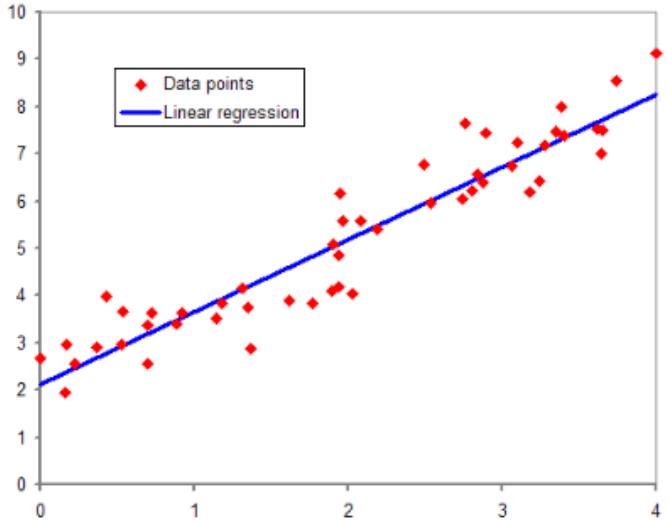
Oh, Hyung Sool

Regression & Classification

- 신경망에 사용되는 회귀식을 **분류(classification)**에 활용하기 위해, 간단히 y를 p로 바꿔보자.
  - **분류(classification)는 범주형 값을 갖는다**

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

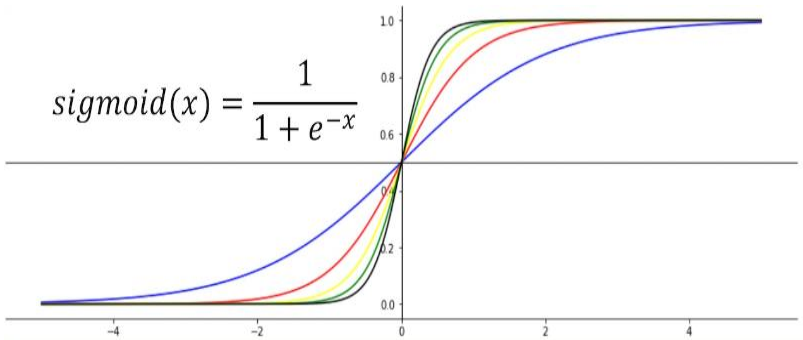
$$p = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$



- p 로 표현한 식을 분류 문제에 적용하기 위해서는 p 값이 제한되어 있지 않다는 문제점을 해결해야만 한다.

- 확률 값으로 만들기 위해 p 값을 0 ~ 1로 제한 시켜야 한다. 그래서 우리가 익히 알고 있는 **시그모이드 함수**를 적용한다.

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

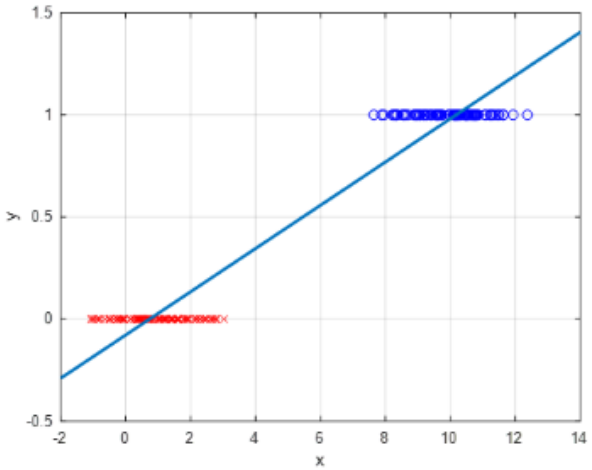


Logistic Regression

- 로지스틱 회귀분석이 필요한 경우
  - ✓ 종속변수가 이항 변수인 경우

    예) 성공/실패, Yes/No, 상승/하락 등

  - ✓ 종속변수가 이항 변수의 특징을 갖는 경우
    - 종속변수의 값이 0 또는 1의 값을 갖는다



- 이항 변수인 경우에 OLS(Ordinary Least Squares) 회귀식을 적용하면?
  - ✓ 회귀식을 해석하는 방법에 문제가 발생
    - 기존의 OLS(Ordinary Least Squares) 회귀분석에서는 종속변수가 연속형 변수에 사용가능함
    - OLS 회귀식이 Y=a+bX 이라면, X가 1 증가할 때, Y가 b만큼 증가?

Sigmoid & Odds

- **선형관계**가 있는 실수의 입력 값들을 토대로 **확률을 예측하는 회귀모델**은 다음과 같다.

$$p = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

- 여기서 문제 발생. 좌변의 확률 값 p는 확률이므로 [0,1] 값만 가능하지만, 우변의 선형 회귀식은 연속형 실수 공간이므로 [-∞,+∞] 값을 갖는다.

- 이 때문에 좌변과 우변 간에 값의 범위가 다른 **미스 매치가 발생**
  그래서, 우변의 [-∞,+∞] 사이의 실수 값을 확률 값으로 변환해줄 수 있는 특별한 식의 필요성이 발생
  이로 인해 등장하는 것이 **승산(Odds)의 개념**이다.

- 승산(Odds)는 실패 확률에 대한 성공 확률의 비이다. Odds=4 란 "성공확률이 실패확률보다 4배 높다" 라는 의미

$$\text{승산(Odds)} = \frac{p}{(1-p)} = \frac{\text{관심 있는 사건이 발생함}}{\text{관심 있는 사건 발생하지 않음}}$$

Odds 의미?

- 경마장 예제
  - ✓ 3마리의 말이 출전하는 상황이며, 3마리 말의 승리 확률은 다음과 같다.
    - A말은 50% 확률로 승리, B말은 30% 확률로 승리, C말은 20% 확률로 승리
    - 경마 참가자는 마권을 1,000원에 사고, 승자를 맞춘 사람에게 3,000을 지불한다면,
    - A말 마권 사는 경우 기대값: 3,000원 × 50% = 1,500원
    - B말 마권 사는 경우 기대값: 3,000원 × 30% = 900원
    - C말 마권 사는 경우 기대값: 3,000원 × 20% = 600원
  - ✓ 그러므로, 경마에 참여하는 모든 사람은 A말 마권을 살 것이다.
  - ✓ 이로 인해, 주최측은 50%의 확률로 A말이 이기면, 판매된 마권 수 × 2,000 만큼 손해
    50%의 확률로 A말이 지면, 판매된 마권 수 × 1,000 만큼 이익
  - ✓ 결론적으로, 주최측은 (1,000원 – 2,000원) × 마권 수 만큼의 손해 발생으로 인해 망하게 된다.

Odds 의미?

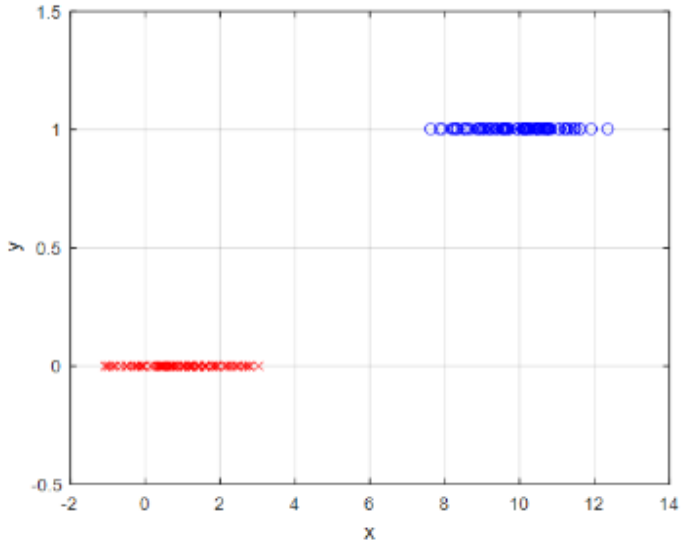
- 경마장 예제: 상금 기준을 변경
  - ✓ 3마리의 말이 출전하는 상황이며, 3마리 말의 승리 확률은 다음과 같다.
    - A말은 50% 확률로 승리, B말은 30% 확률로 승리, C말은 20% 확률로 승리
    - 경마 참가자는 마권을 1,000원에 사고, 승자를 맞춘 사람에게 오즈값의 배로 상금을 지불한다면,
    - A말 마권 사는 경우의 odds= $\frac{p}{1-p} = \frac{0.5}{1-0.5} = 1$, 기대값=1,000원 × 1배 × 50% = 500원
    - B말 마권 사는 경우의 odds=$\frac{p}{1-p} = \frac{0.3}{1-0.3}$ =0.42, 기대값=1,000원 × 0.42배 × 30% = 126원
    - C말 마권 사는 경우의 odds= $\frac{p}{1-p} = \frac{0.2}{1-0.2}$=0.25, 기대값=1,000원 × 0.25배 × 20% = 50원
  - ✓ 그러나, A말의 승리 확률이 3/4=75%라고 한다면?
    - A말 마권 사는 경우의 odds= $\frac{p}{1-p} = \frac{0.75}{1-0.75} = 3$, 기대값=1,000원 × 3배 × 75% = 2,250원
  - ✓ Odds=3 란 "성공확률이 실패확률보다 3배 높다" 라는 의미

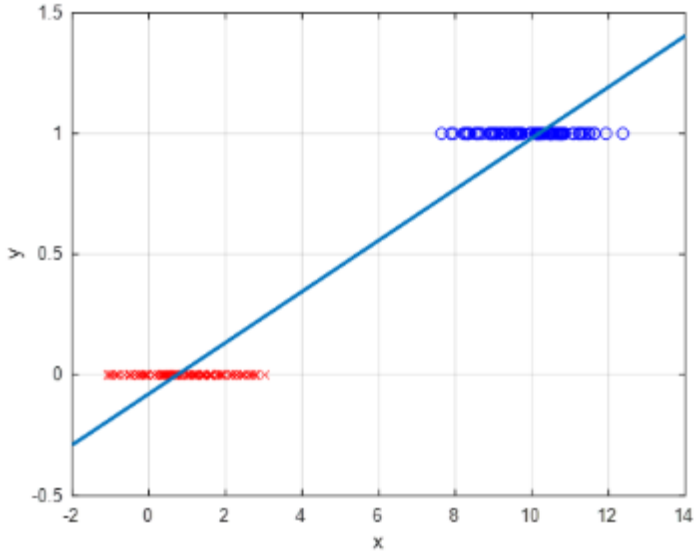
Logistic Regression

- 결과가 범주형인 예제
  라벨이 범주형이라는 것은 가령 "남자, 여자" 혹은 "강아지, 고양이" 등의 연속적인 숫자로 나타내기 어려운 데이터들을 얘기하며, 이 범주들은 보통 0 이나 1로 치환하여 취급한다.

- 아래와 같이 x라는 특성 값이 5보다 작으면 클래스가 0으로, 5보다 같거나 크면 클래스가 1로 결정된다고 생각해 보자

- 범주형 라벨을 가지는 데이터에 대해 기존의 OLS 회귀식을 적용하면 오른쪽과 같은 결과를 얻을 것이다.

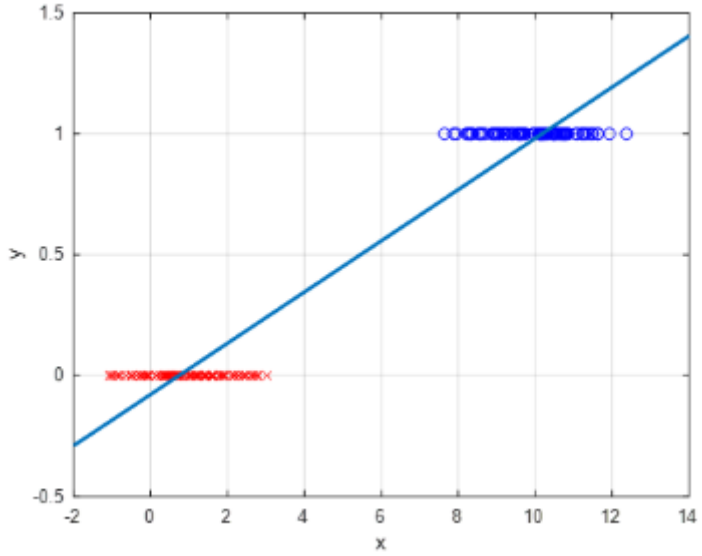


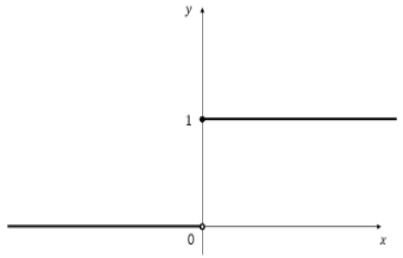


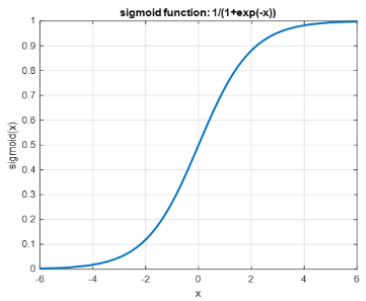
Logistic Regression

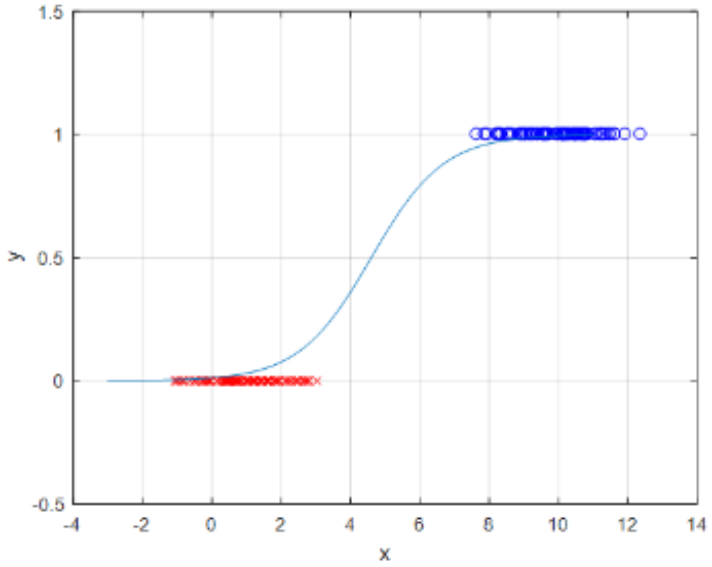
- 범주형 데이터에 대한 모델을 세우기 위해 필요한 함수는 아래의 그림과 같이 어떤 값을 넘어가기 전에는 0, 넘어간 뒤에는 1의 값을 가지는 형태의 함수여야 한다.

- 주어진 데이터에 sigmoid 함수를 적용하면 아래의 그림과 같은 형태일 것이다





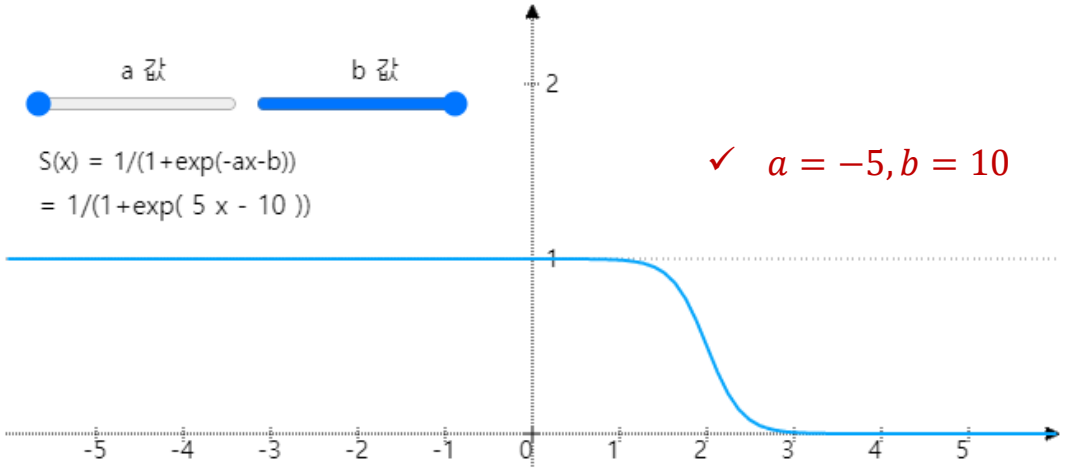




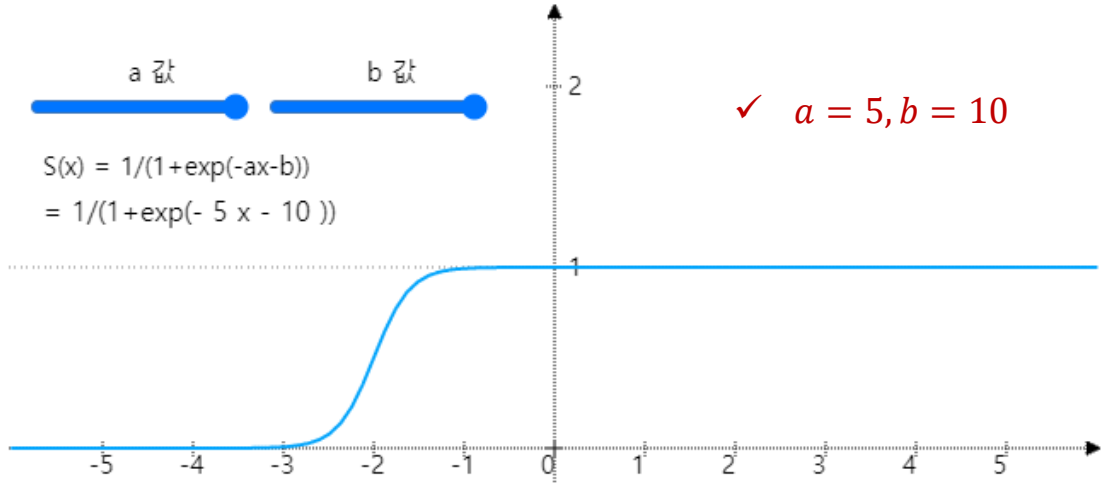
Logistic Regression

- $S(x) = \frac{1}{1+\exp(-ax-b)}$

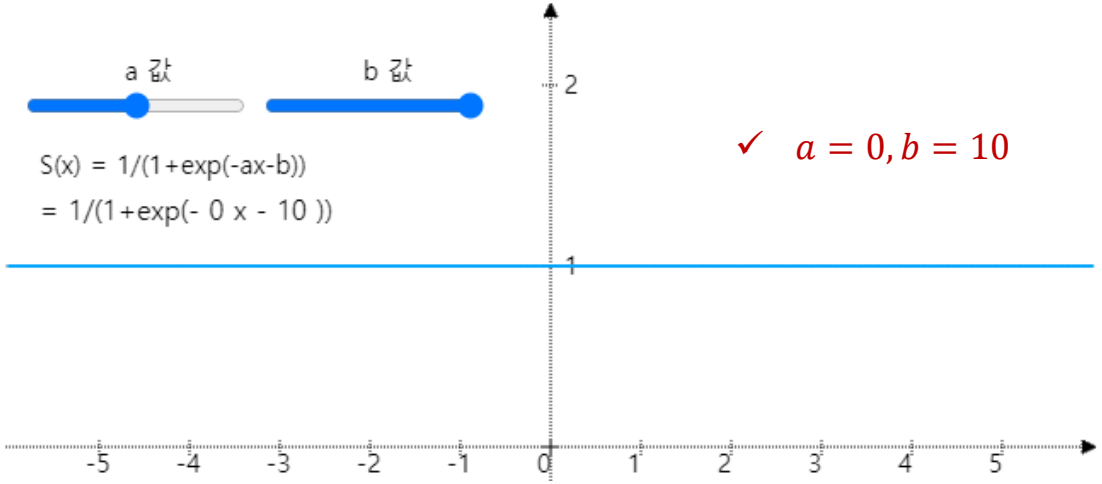
➢ ***기울기인 a가 변화하는 경우***



✓ $a = -5, b = 10$



✓ $a = 5, b = 10$

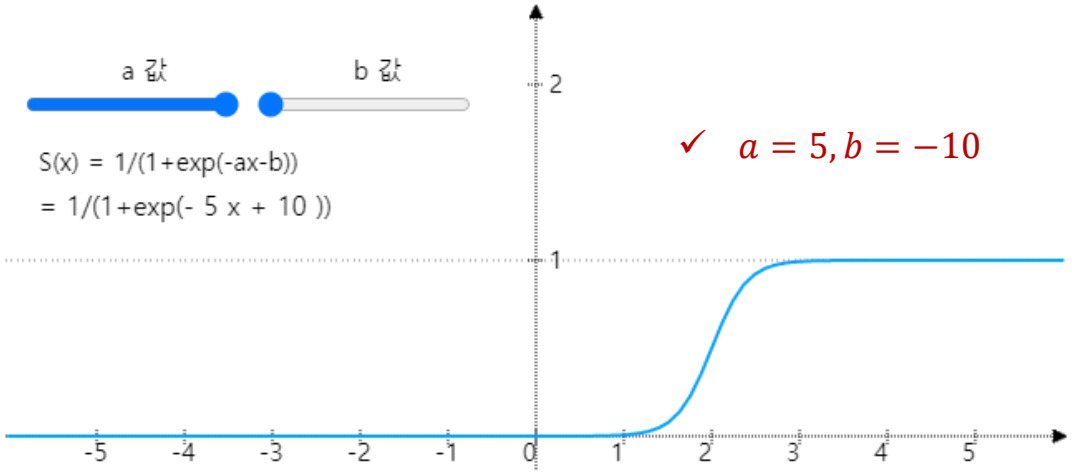


✓ $a = 0, b = 10$

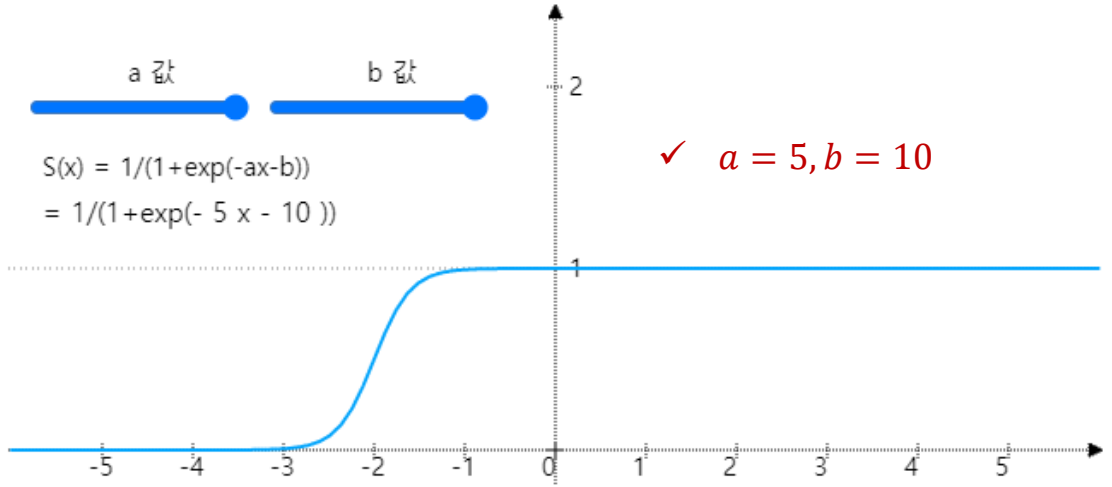
Logistic Regression

- $S(x) = \frac{1}{1+\exp(-ax-b)}$

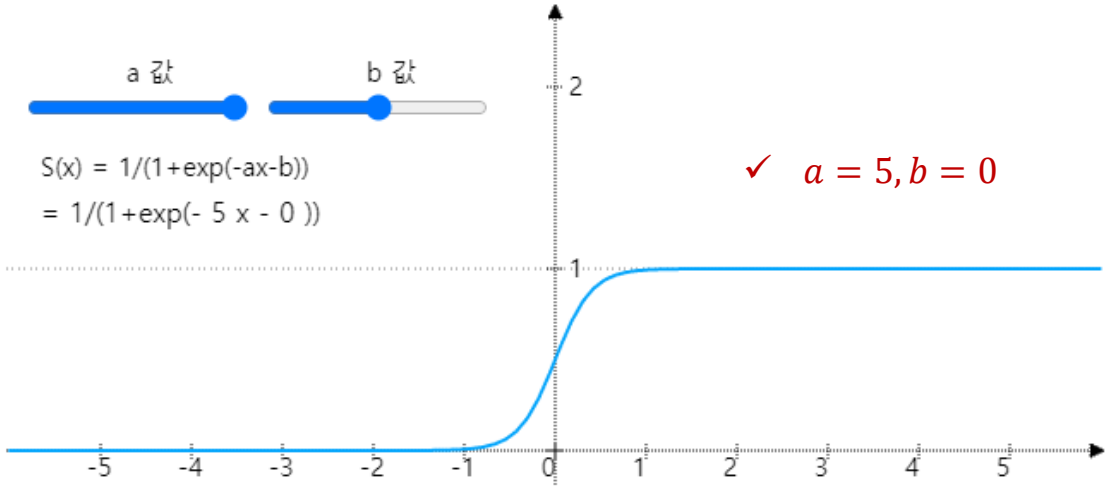
➢ **절편인 b가 *변화하는 경우***



✓ $a = 5, b = -10$



✓ $a = 5, b = 10$



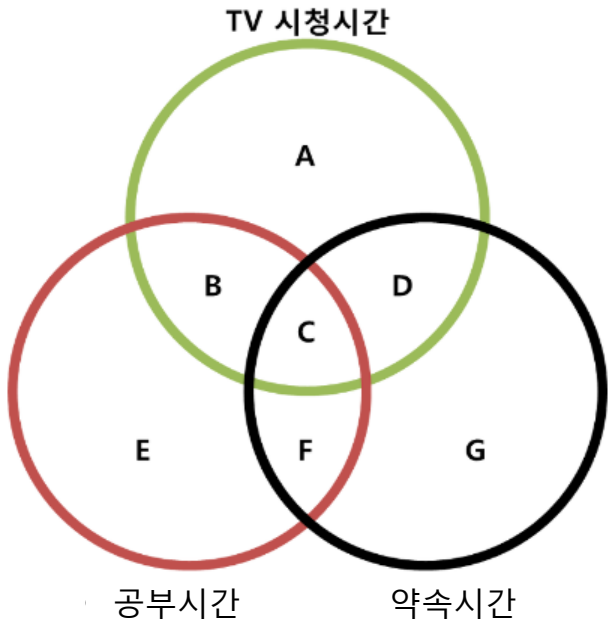
✓ $a = 5, b = 0$

상관계수 의미

- 예제에서 종속변수: TV시청 시간, 독립변수: 공부시간, 약속시간

- 상관계수(zero-order correlation) : B+C , D+C
  - ✓ 0차 상관계수는 계량형인 두 변수간의 상관 정도를 정량화 한 값으로서, Pearson 상관계수 값이다.
  - ✓ 공부시간(E)과 종속변수인 TV시청 시간(A)의 상관관계는 -0.612, 그리고 약속시간(G)과 TV시청 시간의 상관관계는 –0.521이다.

- B+C와 D+C는 각각 공부시간과 약속시간이 Y를 설명하는 정도로서, 0차 상관계수를 제곱한 R^2 은 다음과 같다.
  - 공부시간의 R^2 : (-0.612)^2 = 0.3745
  - 약속시간의 R^2 : (-0.521)^2 = 0.2714
  - ✓ 즉, 공부시간은 Y의 분산을 37.45% 설명하며, 약속시간은 Y의 분산을 27.14% 설명한다.

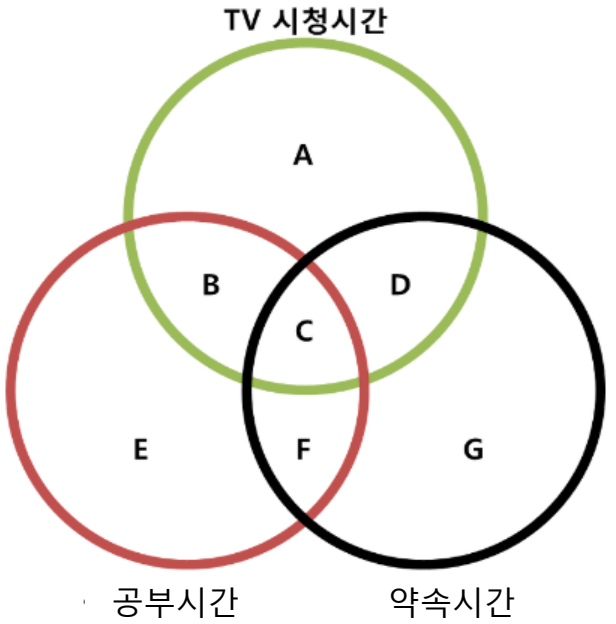


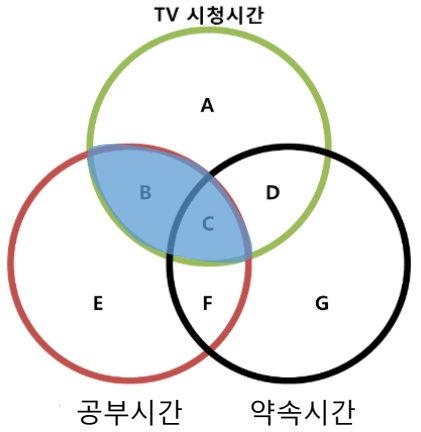
상관계수 의미

- 회귀식의 설명력 (두 변수에 의해 설명되는 Y의 분산) : B+C+D
  - ✓ B+C+D는 공부시간과 약속시간이 결합하여 Y를 설명하는 정도이다.
  - ✓ '공부시간'과 '약속시간'을 투입하여 회귀분석한 결과로 얻어진 설명력 R^2은 0.599(≠0.3745 + 0.2714 = 0.6459) 이다.
  - ✓ 회귀식의 설명력 R^2의 의미는 종속변수의 변동을 어느 정도 설명하는가? 이다.

- 공부시간에 의해 설명되지 않는 TV시청 시간의 분산 : A+D
  - ✓ 공부시간에 의해 설명되지 않는 분산은 1-0.3745 = 0.6255 이다.

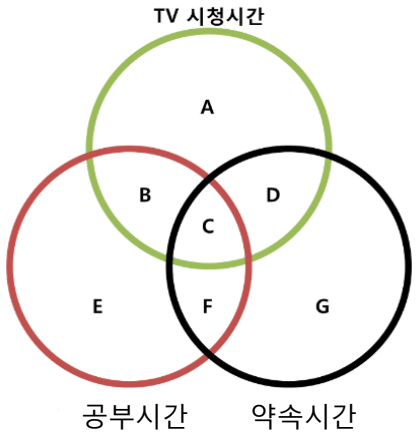
- 편상관계수 (partial correlation) : 약속시간이라는 두번째 독립변수를 고려함으로써 설명되는 부분 : **D**
  - ✓ 편상관계수는 **다른 독립변수의 효과를 제거한 후**(혹은 통제된 상태에서) 한 독립변수(G)와 종속변수(A)의 상관관계이다.



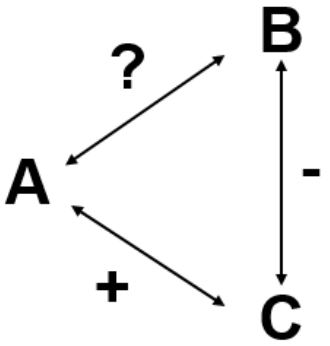


편상관계수 의미

- 편상관분석이 필요한 경우는
  - ✓ TV시청 시간(A)을 종속변수로 하고, '공부시간(E)'과 '약속시간(G)'이 **서로 간에 영향을 주는 변수**임에도 이 두 변수를 서로 다른 독립변수로 **동시에 고려**하는 상황
  - ✓ 이런 상황에서 TV시청 시간(A)에 대한 공부시간(E) 간의 관계를 파악할 때 편상관분석이 필요하다
  - ✓ 즉, 종속변수를 설명하는 독립변수로 고려하는 두 변수(공부시간, 약속시간) 간에도 명백히 강한 상관관계를 갖는 경우에 하나의 독립(설명)변수의 영향력을 제한하고 남은 다른 독립변수**만의** 결과에 대한 영향력을 평가하고자 하는 경우에 분석하는 방법이다.



- 편상관분석을 사용하는 목적
  - ✓ 가짜 상관관계를 찾아내는데 유용하다.
  - ✓ 예: 연봉과 혈압 ~ 나이: 혈압은 나이와 밀접한 상관을 갖는 상황에서, 연봉 결과를 혈압과 나이 독립변수를 동시에 고려하면, 혈압과 나이의 밀접한 상관관계로 인해 연봉과 혈압 간의 상관관계가 있는 것처럼 "가짜 상관관계"의 결과가 나오게 된다.
  - ✓ 그림에서처럼, A-C 간에 강한 + 의 상관관계가 존재하고 동시에 B-C 간에는 강한 – 의 상관관계가 존재하는 경우에, A-B 간에도 상관관계가 있을 것으로 충분히 예상되는 상황에서 상관관계가 없는 격으로 결과가 나올 수 있다. 이와 같은 경우에서도, 편상관분석을 통해 **숨겨진 상관관계**를 찾아낼 수 있다.



다중공선성

- 다중공선성이란?
  - ✓ 회귀분석에서의 다중공선성은 **독립변수 간에 상관관계가 존재하는 것**을 말한다.
  - ✓ 다중공선성이 존재하면 회귀분석의 결과가 왜곡될 수 있다.

- 다중공선성의 원인은 다음과 같다.
  - ✓ 독립변수가 동일한 변수를 다른 방식으로 측정한 경우일 수 있다.

    예) 소득과 소비는 서로 상관관계가 있을 수 있다. 즉, 소득이 증가하면 소비도 증가할 가능성이 높다.

- 다중공선성의 영향은 다음과 같다.
  - ✓ 회귀계수의 표준오차가 증가한다
  - ✓ 회귀계수의 신뢰구간이 넓어진다.
  - ✓ 회귀모델의 예측력이 감소한다.

다중공선성

- 회귀분석 시 독립변수들에 대해서,
  - ✓ 선형성, 정규성, **독립성**, 등분산성을 가정한다.
  - ✓ 회귀분석 모델이 다중공선성 문제를 갖는 경우, 변수 간의 독립성 가정이 위배되면서
    - 회귀모델에 대한 F검정 결과는 유의하게 나타나지만,
    - 개별 독립변수에 대한 통계분석 결과에서 유의하지 않은 변수를 포함하는 경우가 나타난다.
    - 회귀분석 모델의 정확도를 평가할 때 사용하는 대표적인 지표 중 하나가 AIC(Akaike's Information Criterion)
    - AIC는 주어진 데이터 셋에 대한 통계 모델의 상대적인 품질을 평가하는 지표로서, 낮을수록 좋다.

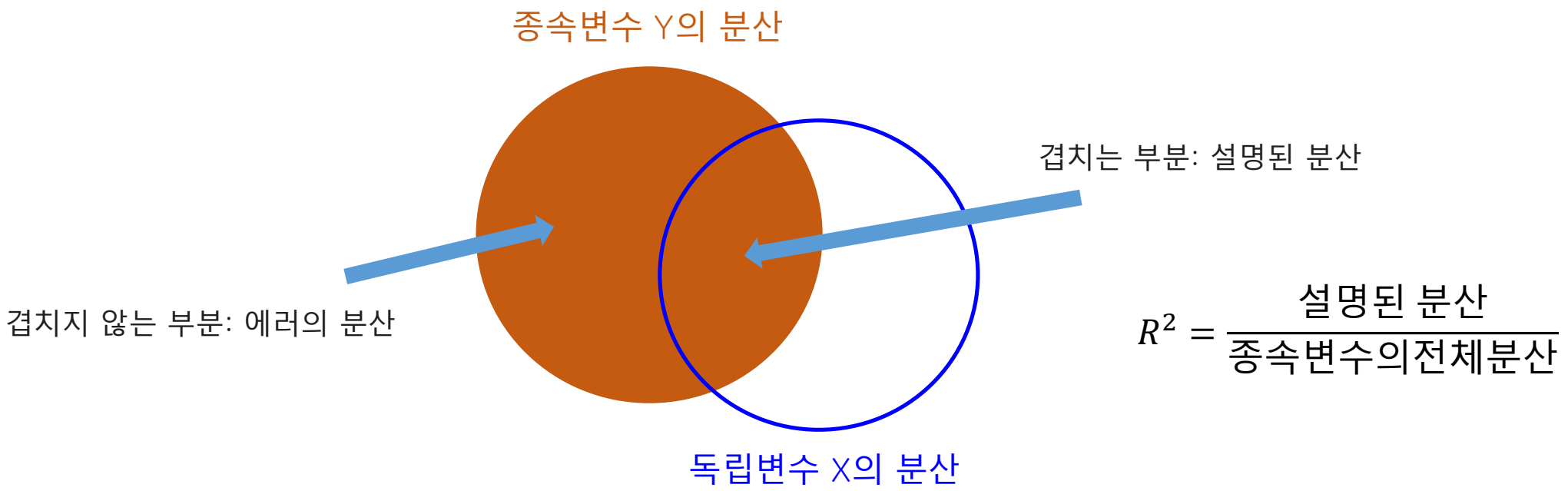
- AIC의 공식은 아래와 같다.

$$AIC = -2\ln(L) + 2k$$

  - ✓ -2ln(L)은 모형의 적합도를 의미하며, L은 Likelihood function 을 의미하고, k는 모형에서 추정되는 파라미터의 개수이다.
  - ✓ 모형의 적합도란 실제 자료와 연구자의 연구 모형이 얼마나 부합하는지를 의미
  - ✓ AIC 값이 작을수록 좋다는 것은 적은 수의 파라미터로 모형의 적합도가 좋은 모델이라는 것을 의미한다

다중공선성

- 회귀분석이란 종속변수의 분산을 독립변수로 설명하는 과정
  - ✓ R^2은 모델의 분산에 대한 설명력이라고 할 수 있다.
  - ✓ R^2이 필요 이상으로 높으면, 과적합(Overfitting)문제가 있을 수 있다.

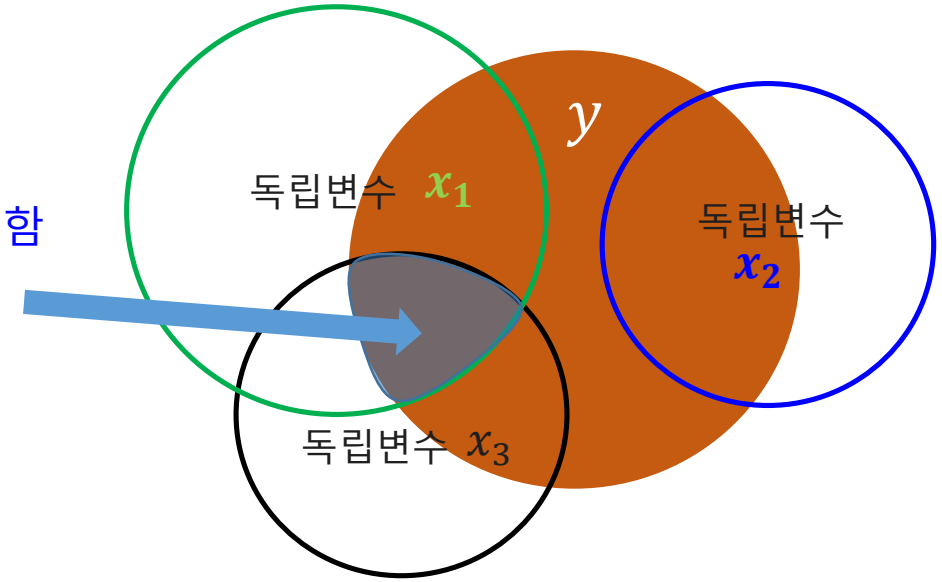


$$R^2 = \frac{\text{설명된 분산}}{\text{종속변수의전체분산}}$$

다중공선성

- 회귀분석의 식이 다음과 같다고 하자.
  - ✓ $y = a + b_1x_1 + b_2x_2 + b_3x_3$
  - ✓ 이때의 회귀분석은 3개의 변수, x_1 , x_2 , x_3 로 y 의 분산(변동)을 얼마나 설명하느냐? 의 문제

x_1과 x_3 간에 다중공선성 문제를 포함



편상관 & 다중공선성

- 다중공선성과 편상관 관계의 차이점
  - ✓ 다중공선성과 편상관 관계 모두 독립변수 간의 관계를 설명하는 개념이지만, 다음과 같은 차이점이 있다.
  - ✓ **다중공선성**은 독립변수 간에 **선형적인** 관계가 존재 한다.

    예) 소득과 소비는 서로 상관관계가 있을 수 있다. 즉, 소득이 증가하면 소비도 증가할 가능성이 높다.
  - ✓ **편상관 관계**는 독립변수 간에 **비선형적인** 관계가 존재 한다.

    예) 교육 수준과 소득은 서로 편상관 관계가 있을 수 있다. 교육 수준이 증가하면 소득도 증가할 가능성이 높지만,
    교육수준이 매우 높아지면 소득이 증가하지 않을 수도 있다.
  - ✓ 다중공선성은 회귀계수의 표준오차를 **증가**시켜 회귀모델의 예측력을 감소시킨다.
  - ✓ 편상관 관계는 회귀계수의 추정을 **왜곡**시켜 회귀모델의 해석을 어렵게 만든다.
  - ✓ 독립변수 간의 상관관계가 높으면 다중공선성이나 편상관 관계가 존재할 가능성이 높다.

- 다중공선성을 확인하는 방법

  ❖ $\mathrm{VIF} = \frac{1}{tolerance} = \frac{1}{1-R^2}$

  - ✓ VIF(variance inflation factor)를 확인. **VIF가 10 이상**이면 다중공선성이 존재할 가능성이 높다.
  - ✓ 공차(tolerance)를 확인. **공차가 0.1 미만**이면 다중공선성이 존재할 가능성이 높다.
  - ✓ 공분산 행렬을 확인. 공분산 행렬의 대각선 요소를 제외한 요소의 절대값이 크면 다중공선성이 존재할 가능성이 높다.

다중공선성 확인 방법

- R에서 VIF 통계량을 구하는 방법
  - ✓ 명령어는 'vif' 이며, 이를 위해서는 'car' 패키지가 필요하다.
  - ✓ install.packages("car")
  - ✓ library(car)

  - ✓ Data set "MASS" 의 "Boston" 데이터를 이용하여
    평균집값(medv) = **f** [지역의 소득수준(lstat), 방의 수(rm), 선생 한명이 담당하는 학생수(ptratio)]
    에 대한 회귀분석을 하였고
  - ✓ 이 회귀분석에서 3개 독립변수 들간에 다중공선성이 없는가? 를 확인하였음
  - ✓ "vif' 분석 결과 값이 1.679**, 1.653**, 1.198** 로서, 10보다 충분히 작기 때문에 다중공선성은 없다고 할 수 있다.

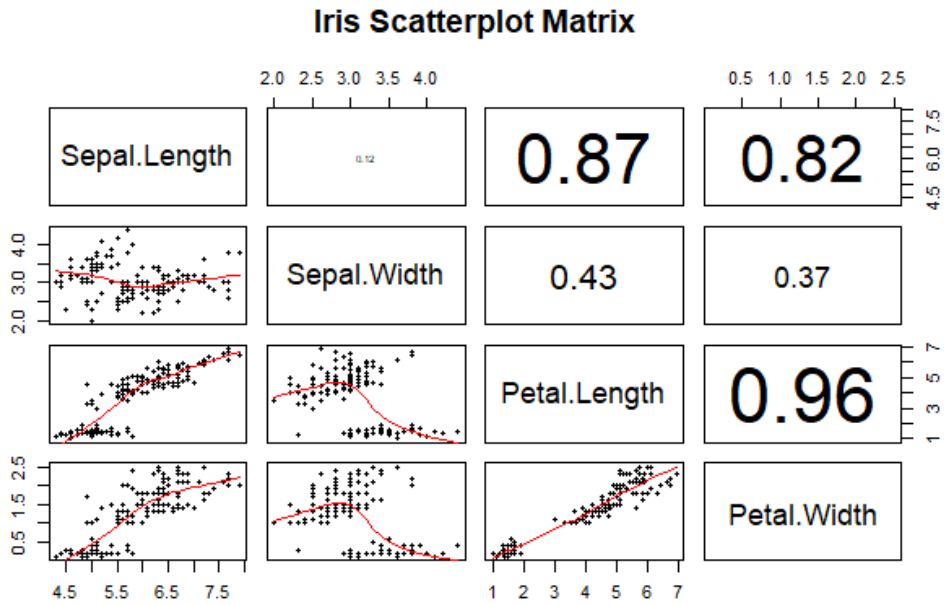
```
25  install.packages("car")
26  library(car)
27
28  model <- lm(medv ~ lstat + rm + ptratio, data = Boston)
29  summary(model)
30  vif(model)
```

```
> library(MASS)
> data()
> model <- lm(medv ~ lstat + rm + ptratio, data = Boston)
> summary(model)

Call:
lm(formula = medv ~ lstat + rm + ptratio, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4871  -3.1047  -0.7976   1.8129  29.6559

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
rm           4.51542    0.42587  10.603  < 2e-16 ***
ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.229 on 502 degrees of freedom
Multiple R-squared:  0.6786,    Adjusted R-squared:  0.6767
F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
> vif(model)
   lstat       rm  ptratio
1.679425 1.653419 1.198101
```

편상관 검정 방법

- 편상관 관계를 시각적으로 확인하는 방법: **산점도 행렬**(scatterplot matrix)
  - ✓ 산점도 행렬은 독립변수 간의 관계를 한 번에 시각화할 수 있는 방법
  - ✓ 산점도 행렬을 통해 편상관 관계를 확인하는 방법은 다음과 같다.
    - 독립변수 간에 선형적인 관계가 아닌 비선형적인 관계가 나타나는 경우, 편상관 관계가 존재할 가능성이 높다
    - 독립변수 간에 과도하게 밀집된 영역이 나타나는 경우, 편상관 관계가 존재할 가능성이 높다.




```
pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data=iris
      lower.panel=panel.smooth, upper.panel=panel.cor,  pch=20,
      main="Iris Scatterplot Matrix")
```

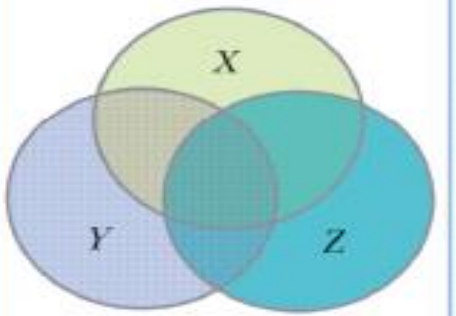
편상관 계수 및 검정 방법

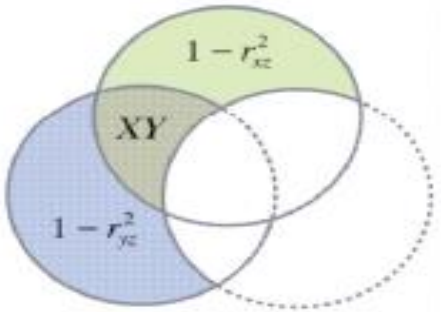
편상관계수

- 제3의 변수가 두 변수에 미치는 영향을 제거한 후, 두 변수 간의 순수한 상관관계를 나타내는 계수임
- 제3의 통제변수를 Z라 할 때, 두 변수 X와 Y간의 편상관계수는 Z가 X와 Y에 미치는 선형효과(linear effect)를 제거시킨 뒤 남은 잔차 간의 상관계수를 의미함

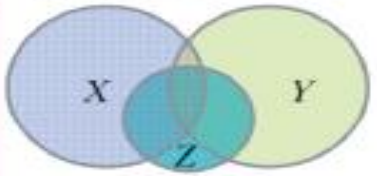
$$r_{xy \cdot z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\ \sqrt{1 - r_{yz}^2}}$$

세 변수 간의 선형관계 → 한 변수의 영향력을 제외시킴 → 두 변수 간의 순수한 선형관계

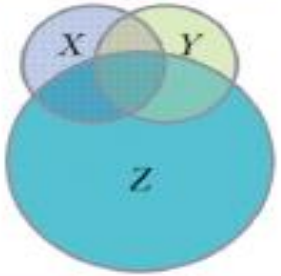




- 변수 Z가 미치는 영향의 특성에 따라 편상관계수는 일반상관 계수보다 작거나 클 수 있음



$r_{xy \cdot z} < r_{xy}$



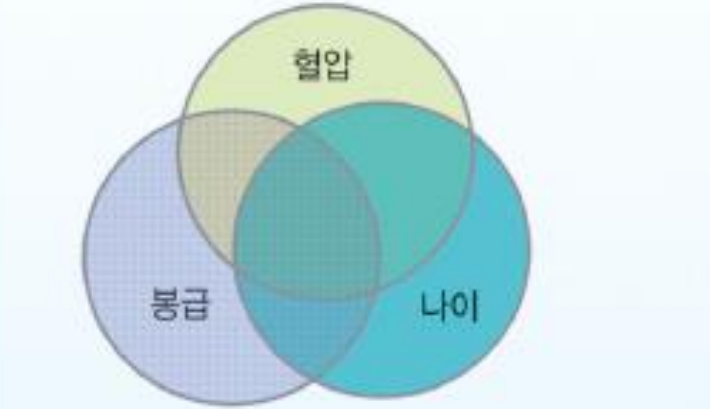
$r_{xy \cdot z} > r_{xy}$

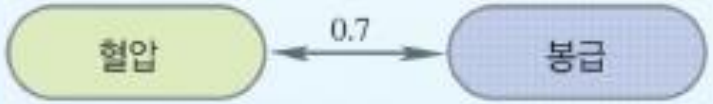
편상관 계수 및 검정 방법

- 혈압과 봉급 간의 상관계수를 구하고 나이가 혈압과 봉급에 미치는 영향을 제거한 후, 혈압과 봉급 간의 순수한 편상관계수를 구하여 비교설명함

혈압과 봉급의 상관계수

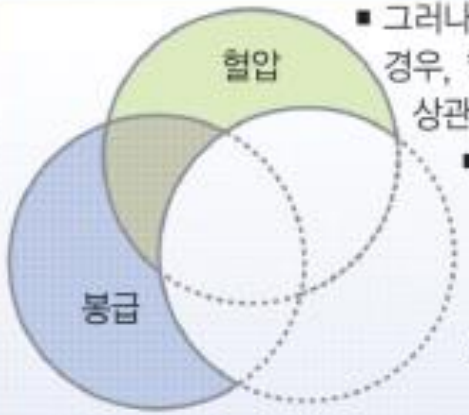
- 혈압과 봉급 간에는 높은 상관관계가 있음

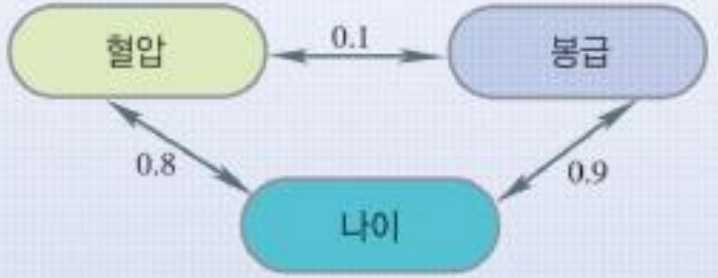




혈압과 봉급의 편상관계수(나이 변수의 영향력 제외)

- 그러나 나이를 통제했을 경우, 혈압과 봉급의 상관관계는 매우 낮음
- 혈압과 봉급 간의 상관관계는 대부분 나이로 인하여 비롯되고 있음을 알 수 있음





편상관 계수 및 검정 방법

```
##### mtcars 데이터셋을 이용한 편상관분석 예제  #####

head(mtcars)
str(mtcars)

### mpg:연비, cyl:실린더 수, hp: 마력, wt: 차량무게
### 4개 변수(특성)로 데이터 재구성
mt <- mtcars[,c("mpg","cyl","hp","wt")]

### 4개 변수들간의 상관계수 계산
cor(mt)

round(cor(mt),2)

### 편상관계수와 검정을 위한 패키지를 설치
install.packages("ppcor")
library(ppcor)

### 편상관 계수 및 검정 실시
pcor(mt)
```

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> mt <- mtcars[,c("mpg","cyl","hp","wt")]
> cor(mt)
           mpg        cyl         hp         wt
mpg  1.0000000 -0.8521620 -0.7761684 -0.8676594
cyl -0.8521620  1.0000000  0.8324475  0.7824958
hp  -0.7761684  0.8324475  1.0000000  0.6587479
wt  -0.8676594  0.7824958  0.6587479  1.0000000
> round(cor(mt),2)
      mpg   cyl    hp    wt
mpg  1.00 -0.85 -0.78 -0.87
cyl -0.85  1.00  0.83  0.78
hp  -0.78  0.83  1.00  0.66
wt  -0.87  0.78  0.66  1.00
> install.packages("ppcor")
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ppcor_1.1.zip'
Content type 'application/zip' length 30012 bytes (29 KB)
downloaded 29 KB

패키지 'ppcor'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다

다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
        C:\Users\User\AppData\Local\Temp\Rtmp8wqG1p\downloaded_packages
```

편상관 계수 및 검정 방법

```
> cor(mt)
            mpg        cyl         hp         wt
mpg   1.0000000 -0.8521620 -0.7761684 -0.8676594
cyl  -0.8521620  1.0000000  0.8324475  0.7824958
hp   -0.7761684  0.8324475  1.0000000  0.6587479
wt   -0.8676594  0.7824958  0.6587479  1.0000000
```

- 대각선의 상관계수 값 1.00000은 자신과의 상관관계로서 보는 것처럼 당연히 1이 된다.

```
> cor(mt)
            mpg        cyl         hp         wt
mpg   1.0000000 -0.8521620 -0.7761684 -0.8676594
cyl  -0.8521620  1.0000000  0.8324475  0.7824958
hp   -0.7761684  0.8324475  1.0000000  0.6587479
wt   -0.8676594  0.7824958  0.6587479  1.0000000
```

- mpg 와 cyl 간의 상관계수 값은 -0.8521620 이며, 이 의미는 cyl 값이 커질수록 mpg 는 줄어든다는 의미
- hp 와 wt 간의 상관계수 값은 0.6587479 이며, 이 의미는 hp값이 커질수록 wt 가 커진다는 의미
- 하지만, 현재의 상관계수 값은 하나의 독립변수에 다른 2개 독립변수의 영향이 동시에 **중복, 반영된 것 일수 있다**. 따라서, 다른 2개 독립변수의 영향을 배제한 **편상관계수에 대한 분석**이 필요

편상관 계수 및 검정 방법

```
> library(ppcor)
> pcor(mt)
$estimate
           mpg        cyl         hp         wt
mpg  1.0000000 -0.3073687 -0.2758932 -0.6285559
cyl -0.3073687  1.0000000  0.5340905  0.2224468
hp  -0.2758932  0.5340905  1.0000000 -0.1574640
wt  -0.6285559  0.2224468 -0.1574640  1.0000000

$p.value
             mpg         cyl          hp           wt
mpg 0.0000000000 0.098480097 0.140015155 0.0001994765
cyl 0.0984800975 0.000000000 0.002365994 0.2374063384
hp  0.1400151550 0.002365994 0.000000000 0.4059618058
wt  0.0001994765 0.237406338 0.405961806 0.0000000000

$statistic
           mpg       cyl        hp        wt
mpg  0.000000 -1.709183 -1.518838 -4.276365
cyl -1.709183  0.000000  3.342856  1.207328
hp  -1.518838  3.342856  0.000000 -0.843747
wt  -4.276365  1.207328 -0.843747  0.000000

$n
[1] 32

$gp
[1] 2

$method
[1] "pearson"
```

```
> cor(mt)
            mpg        cyl         hp        wt
mpg  1.0000000 -0.8521620 -0.7761684 -0.8676594
cyl -0.8521620  1.0000000  0.8324475  0.7824958
hp  -0.7761684  0.8324475  1.0000000  0.6587479
wt  -0.8676594  0.7824958  0.6587479  1.0000000
```

- "pcor(mt)"에 의한 편상관분석 결과를 상관분석 결과와 비교해보면,

  ✓ : -0.8521620 -> -0.3073687 로 hp , wt 변수의 영향을 배제한 mpg 와 cyl 만을 고려한 편상관계수 값이 줄어든 것을 확인할 수 있다.

  ✓ : mpg 와 cyl 간의 편상관분석을 통한 상관계수 -0.307 에 대한 t검정 결과에 대한 p-value 값이 0.09848로서, 유의수준 5%(0.05) 보다 크기 때문에 귀무가설이 채택.

즉, hp와 wt 영향을 배제하고 mpg 와 cyl 두 변수만을 고려하였을 때, mpg 와 cyl 변수 간의 상관계수 값이 -0.30733687 로 계산되었지만 **mpg 와 cyl 변수 간에는 상관관계가 있다고 할 수 없다**는 결론을 얻게 된다.