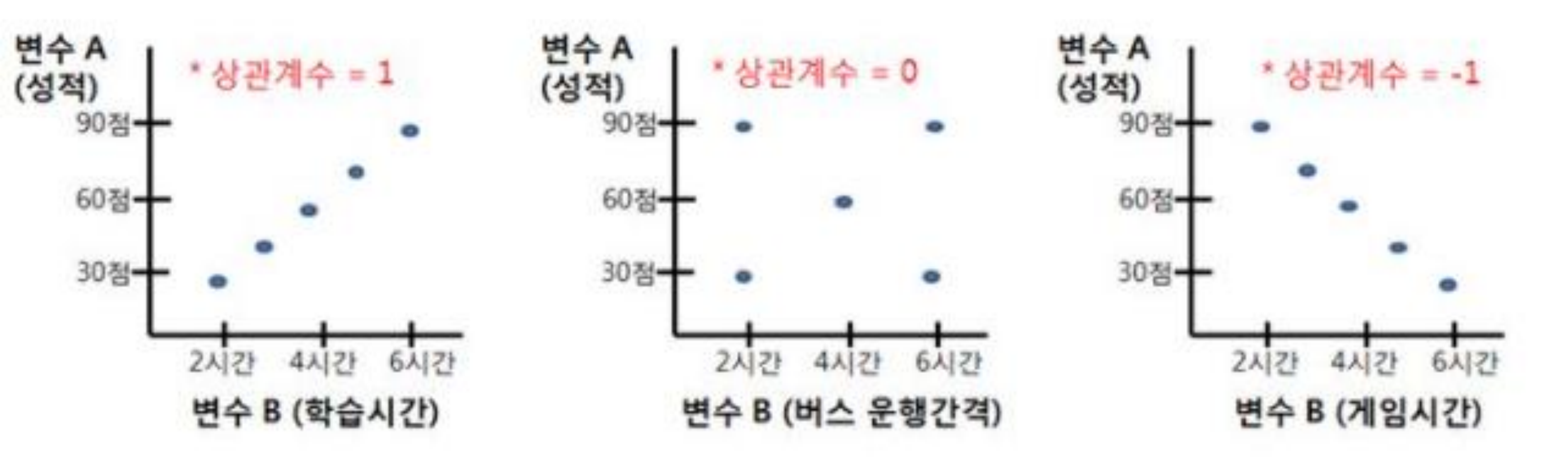
상관분석 개념(정의)

- 연속 변수로 측정된 두 변수간의 선형 관계를 분석하는 기법
- x가 증가함에 따라 y도 증가(감소)되는지를 분석



〈출처: https://sooupforlee.tistory.com/entry/SPSS-%EB%A6%AC%EC%84%9C%EC%B9%98-11-%EC%83%81%EA%B4%80%EA%B4%80%EA%B3%84-%EB%B6%84%EC%84%9D-correlation

기본 가정사항

1) 두 변수 중 적어도 하나의 변수는 정규분포일 것

- 정규성 검사: shapiro.test()

*** 만약, 두 변수 모두 정규성을 만족하지 못한다.**

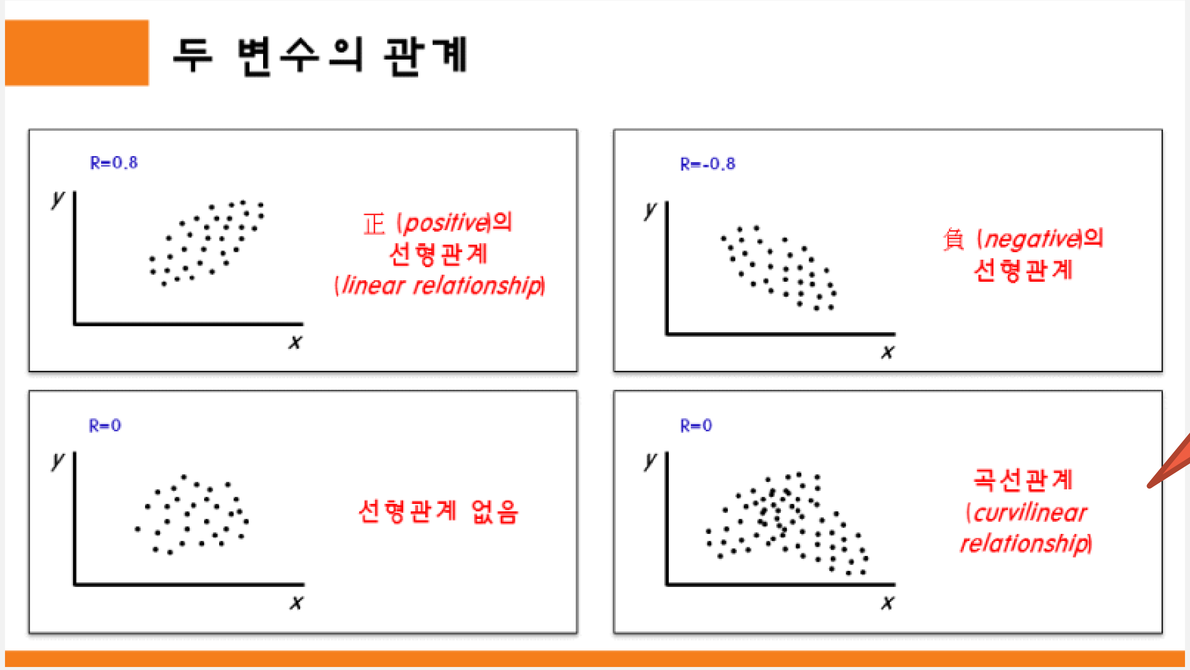
Spearman, Kendall 상관계수

=〉 정규성 검정에서 정규분포를 따르지 않거나 표본의 개수가 10개 미만일 때 사용

기본 가정사항

2) 연속형 두 변수 간에는 선형적인 관계일 것

(분석을 실시하기 전, 반드시 두 변수간의 산점도를 통해 확인!)



아래의 두개 표는
피어슨 상관분석
시행 불가!!!

〈출처: https://m.blog.naver.com/PostView.nhn?blogId=y4769&logNo=220227007641&proxyReferer=https:%2F%2Fwww.google.com%2F

공분산(Covariance)

○ 2개의 확률 변수의 상관 정도를 나타내는 값

○ 만약 하나의 값이 상승하는 경향을 보이면서 다른 값도 상승

➔ 공분산 값은 양수, 반대면 음수를 보임

○ 공분산 값만으로는 상승, 하강 경향을 알 수는 있으나
어느정도의 상관관계인지는 알 수 없음

➔ 따라서 공분산을 표준화 시킨 "상관계수 " 를 통해 파악!

〈출처: https://ordo.tistory.com/21〉

상관관계와 피어슨 상관계수(Pearson Correlation Coefficient)

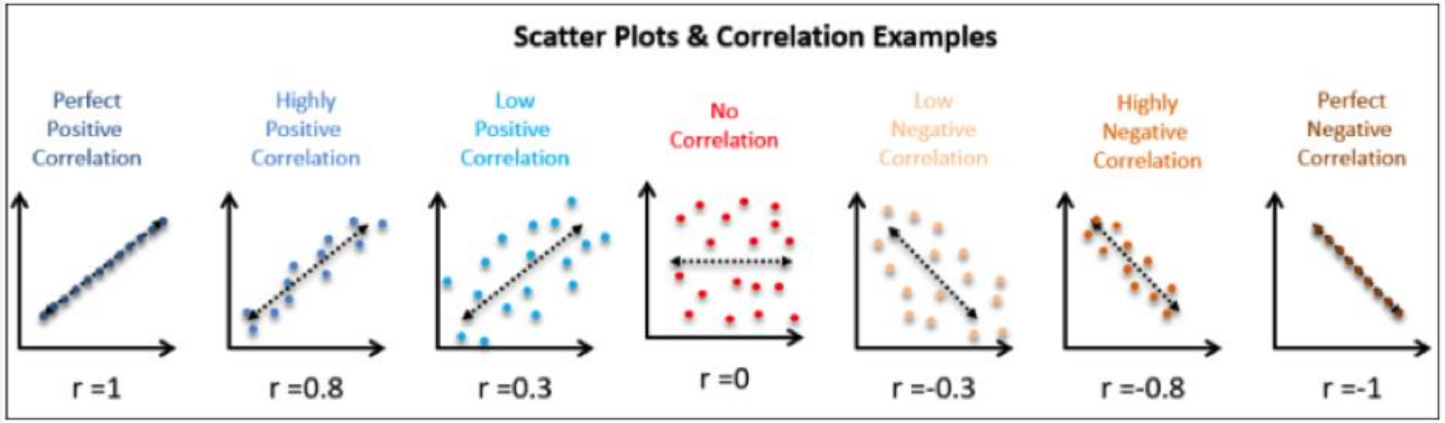
- 두 변수의 선형적인 관계 정도를 나타냄
- 일반적으로, 피어슨 상관계수를 의미
- 피어슨 상관계수 공식

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- x_i, y_i : 표본 집단의 x, y값
- $\bar{x}$, $\bar{y}$: x, y의 값에대한 평균

〈출처: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient〉

상관관계와 피어슨 상관계수(Pearson Correlation Coefficient)



r이 -1.0과 -0.7 사이이면, **강한 음적 선형관계**

r이 -0.7과 -0.3 사이이면, **뚜렷한 음적 선형관계**

r이 -0.3과 -0.1 사이이면, **약한 음적 선형관계**

r이 -0.1과 +0.1 사이이면, **거의 무시될 수 있는 선형관계**

r이 +0.1과 +0.3 사이이면, **약한 양적 선형관계**

r이 +0.3과 +0.7 사이이면, **뚜렷한 양적 선형관계**

r이 +0.7과 +1.0 사이이면, **강한 양적 선형관계**

〈출처: https://astinaccounts.com/easy-accounting-how-to-use-excel-for-data-analytics/〉

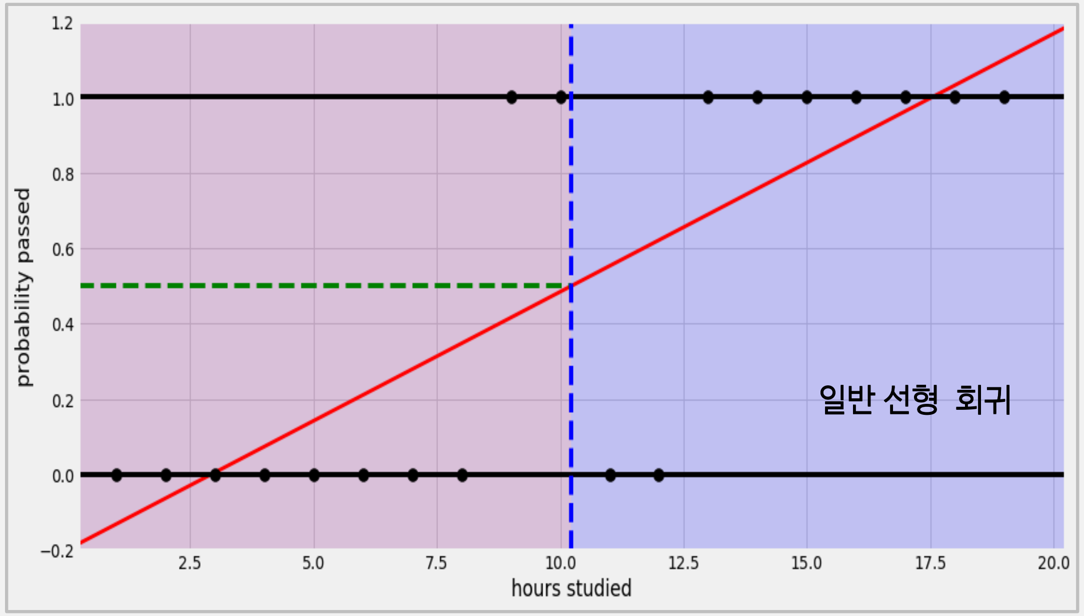
로지스틱 회귀 분석 ?

- 동전을 던져서 결과가 앞인가?
- 메일을 수신받았는데 메일이 스팸메일인가?
- 환자의 상태가 종양인가? 암인가?

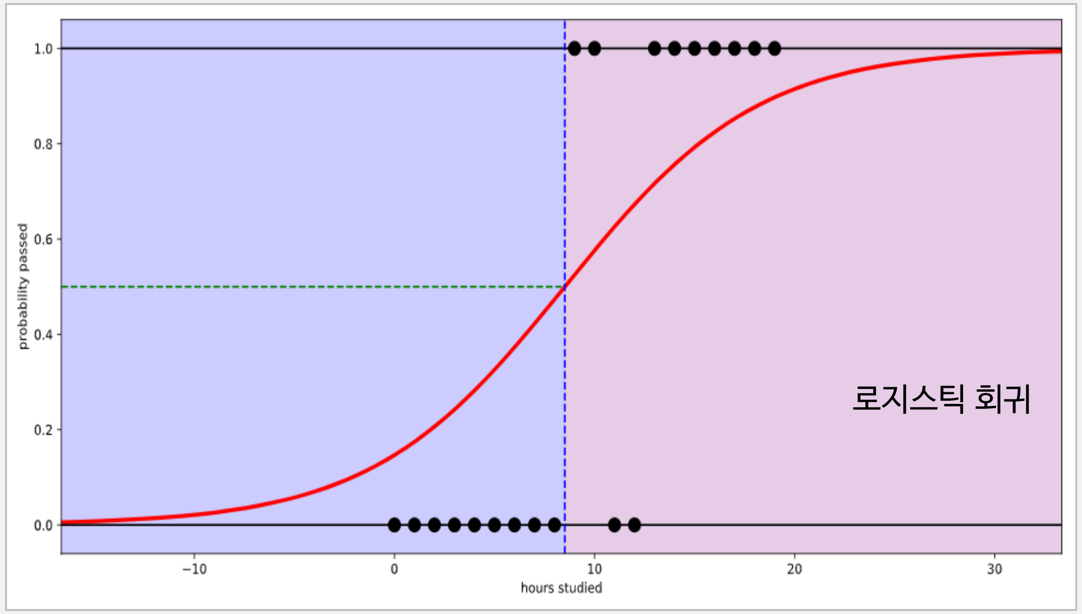
로지스틱 회귀 분석 ?

- **로지스틱 회귀(Logistic Regression)**는 회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도 학습 알고리즘

※ 즉, 로지스틱 회귀는 1,0 즉 (이다 /아니다)로 구별



VS



〈참고: http://hleecaster.com/ml-logistic-regression-concept/〉

로지스틱 회귀 목적 ?

- 독립변수와 종속변수의 관계를 찾음으로써, 새로운 독립변수의 집합이 주어졌을때, 종속변수의 값을 예측할 수 있음

이항 로지스틱 회귀와 다항 로지스틱 회귀

- 이항 로지스틱 회귀

  - 범주가 두개인 결과 변수 예측

- 다항 로지스틱 회귀

  - 2개보다 많은 결과 변수 예측

로지스틱 회귀 3가지 요소

1. Odds
2. Logit 변환
3. 시그모이드 함수

1. Odds(승산비) ?

- 범주 0에 속할 확률 대비 범주 1에 속할 확률

EX) 게 임 아이템을 강화를 한다.

(성공확률 80%, 실패확률 20%)

$$\text{odds} = \frac{p(y=1|x)}{1-p(y=1|x)}$$

승산비 = $\frac{80\%(\text{성공확률})}{20\%(\text{실패확률})}$

※ 1보다 크다는 것은 예측변수가 증가하면 결과가 발생할 승산도 증가한다

2. Logit 변환

- odds에 log를 앞에 붙인 형태를 Logit 변환이라고 함

$$\text{logit}(p) = \log \frac{p}{1-p}$$

- Log를 붙이면 형태가 선형형태로 바뀌고 수식도 간단해짐

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

↓

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

3. 시그모이드 함수

- 확률을 0에서 1사이로 커브 모양으로 나타내야 하는데, 이걸 가능하게 해주는 게 바로 Sigmoid 함수다.
- 시그모이드 함수는 결과 값을 0~1 사이의 값으로 변환해주는 역할만 한다
- Odds를 Sigmoid 함수에 넣어서 0~1사이 값으로 변환해준다

$$\text{logistic function} = \frac{e^{\beta.X_i}}{1 + e^{\beta.X_i}}$$

〈참고: https://shurain.net/personal-perspective/logistic-regression/ 〉

로그 가능도(Log-likelihood)

- 가능도 - 가정된 분포에서 주어진 데이터가 나올 확률
- 계산과 편의를 위해 일반적으로 가능도함수에 로그함수를 씌어 사용
- GLM은 최소제곱법이 아닌 최대가능도추정법을 이용

$$log - likelihood = \sum_{i=1}^{N}\left[Y_i ln\left(P(Y_i)\right) + (1 - Y_i) ln\left(1 - P(Y_i)\right)\right]$$

이탈도

- 로지스틱 회귀모형이 얼마나 데이터를 못 설명하는지에 대한 척도
- 어떤 모형의 a의 최대로그우도에서 포화모형 b의 최대로그우도를 뺀것에 -2를 곱한것
- 카이제곱분포를 사용하기 때문에 이탈도 값의 유의성을 계산하기 쉽기 때문에 로그가능도 보다 더 많이 사용을 한다
- 이탈도가 낮을 수록 좋은 모형임

$$\text{이탈도} = -2[\log(L_m) - \log(L_s)]$$

AIC

- AIC는 입력변수의 수가 증가한다고 항상 작아지지는 않으므로 가장 작은 AIC를 가지는 모형을 선택
- AIC 값은 낮을 수록 좋다
- L_M은 모형 M에 대한 우도함수의 최대값, p는 모수의 수이다.

$$AIC = -2\log(L_M) + 2p$$

- L_M 은 모형 M에 대한 우도함수(likelihood function)의 최대값 p는 모수의 수

〈참고: http://www.jangun.com/study/DataMining.html/ 〉

참고자료

< 상관분석 참고자료 >

- https://m.blog.naver.com/PostView.nhn?blogId=y4769&logNo=220227007641&proxyReferer=https:%2F%2Fwww.google.com%2F
- https://kim-mj.tistory.com/56

<로지스틱 회귀 참고자료>

- https://nittaku.tistory.com/478
- https://wikidocs.net/34034
- https://www.rdocumentation.org/packages/mlogit/versions/1.1-1/topics/mlogit
- https://m.blog.naver.com/y4769/221851780608