고고학 자료 통계분석

Week 7 : 회귀분석

숭실대학교 사학과 석사과정 1학기
주 찬 혁

계획

주차	제목	내용
1	Intro	소개, R 설치
2	기초 통계(1)	모집단과 표본, 기술통계량
3	기초 통계(2)	변수의 종류, 가설과 검정, 오류, 분석절차
4	전처리	데이터 전처리
5	시각화	다양한 종류의 그래프
6	검정과 상관분석	검정, 상관분석
7	회귀분석	선형회귀, 다중선형회귀, 로지스틱회귀
8	군집분석	K-means,
9	판별분석	DA, MDA
10	주성분분석	PCA

복습

- 검정이 무엇인지에 대해 안다.
- R로 다양한 검정을 시행할 수 있다.
- Median Polish가 무엇인지 안다.
- R로 Median Polish를 시행할 수 있다.
- 상관분석에 대해 안다.
- R로 상관분석을 시행할 수 있다.

회귀분석이란?

- 둘 이상의 **변수 사이의 관계를 확인**하고 적합도를 측정하는 분석
- 각 변수 사이의 **인과관계**를 추정할 수 있음
- 회귀식을 통해 **예측**을 할 수 있음
- 데이터의 형태가 **선형인지 비선형인지** 파악
- **결측치 및 이상치**에 대한 처리 필요

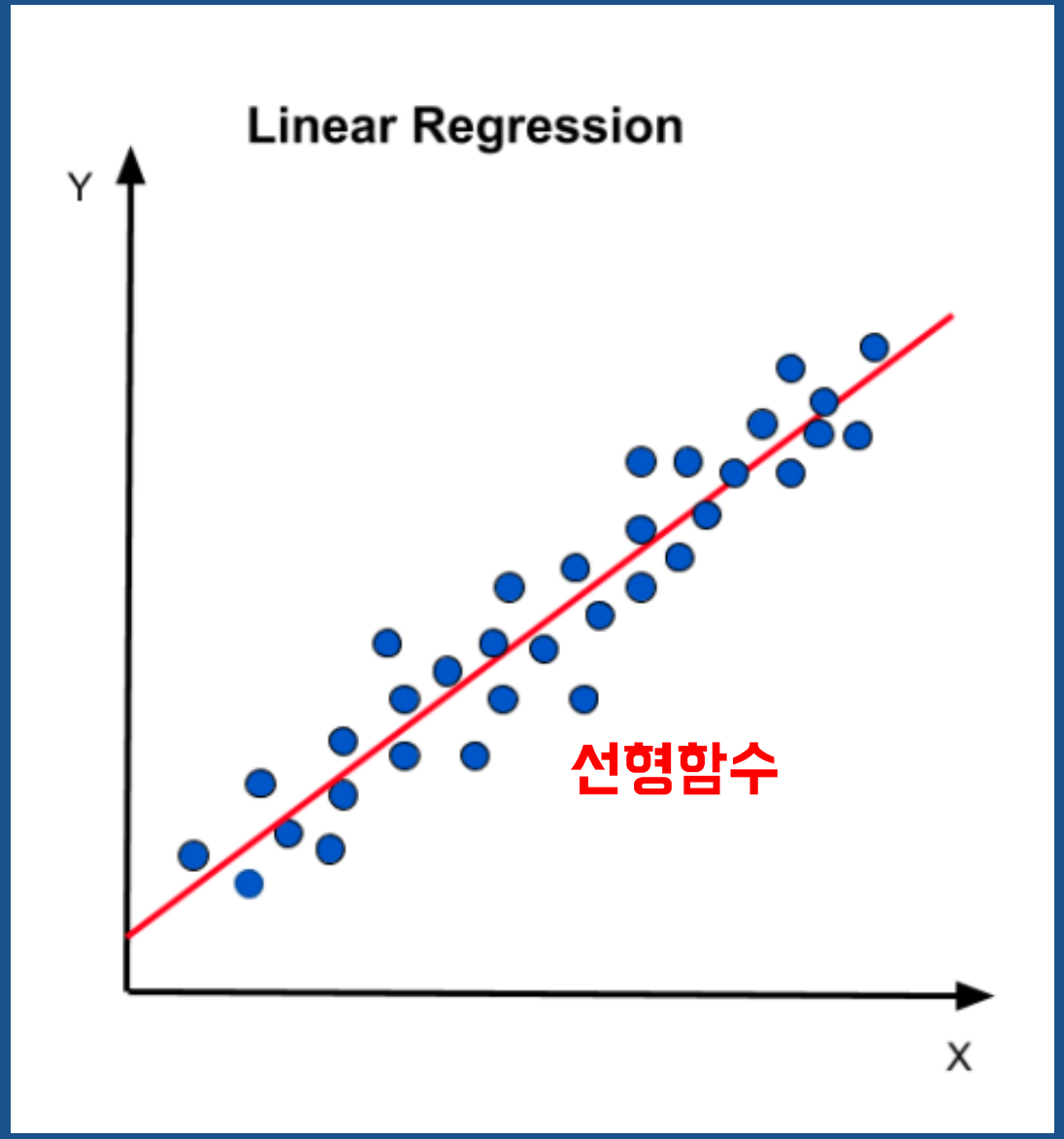
회귀분석의 종류

- 선형 회귀
    - 단순선형 회귀
    - 다중선형 회귀
- 로지스틱 회귀
- 리지 회귀
- 라쏘 회귀

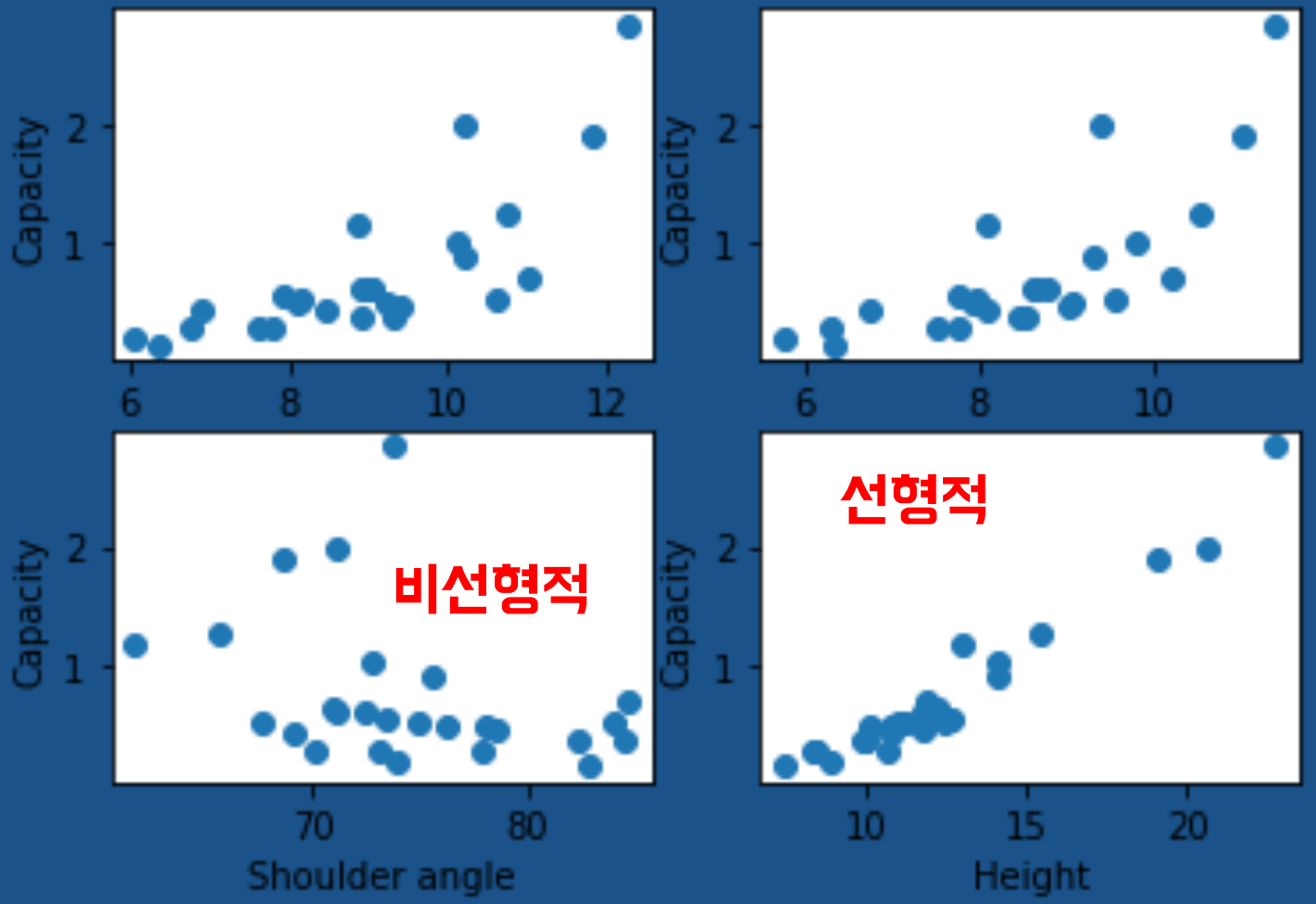
...

선형성이란?

- **직선적으로 똑바른 성질**
- 잔차분석을 통해 검정
- $y = ax + b$
  - a : 회귀계수
  - b : 상수



선형성이란?


Capacity
2
1
6
8
10
12
Capacity
2
1
6
8
10
비선형적
Capacity
2
1
70
80
Shoulder angle
선형적
Capacity
2
1
10
15
20
Height

단순선형 회귀분석

- 하나의 독립변수(x)가 하나의 종속변수(y)에 미치는 영향을 분석하는 방법
- **선형적 상관관계**를 모델링하여 분석하는 회귀분석의 일종.
- 장점
  - 각 값들이 시사하는 것이 명확하여 **해석하기 용이**함
  - 분석 속도가 빠름
- 단점
  - 반드시 두 변수의 관계가 **선형적이라는 가정** 위에서 시행됨
  - 이상치에 민감함

단순선형 회귀분석

> 모델명 <- lm(y ~ x, 데이터)

> summary(모델명)

- lm() 명령어를 사용하여 회귀분석 실행
- summary를 통해 모델의 정보 출력
- 확인된 회귀계수와 상수를 통해 회귀식 산출 가능

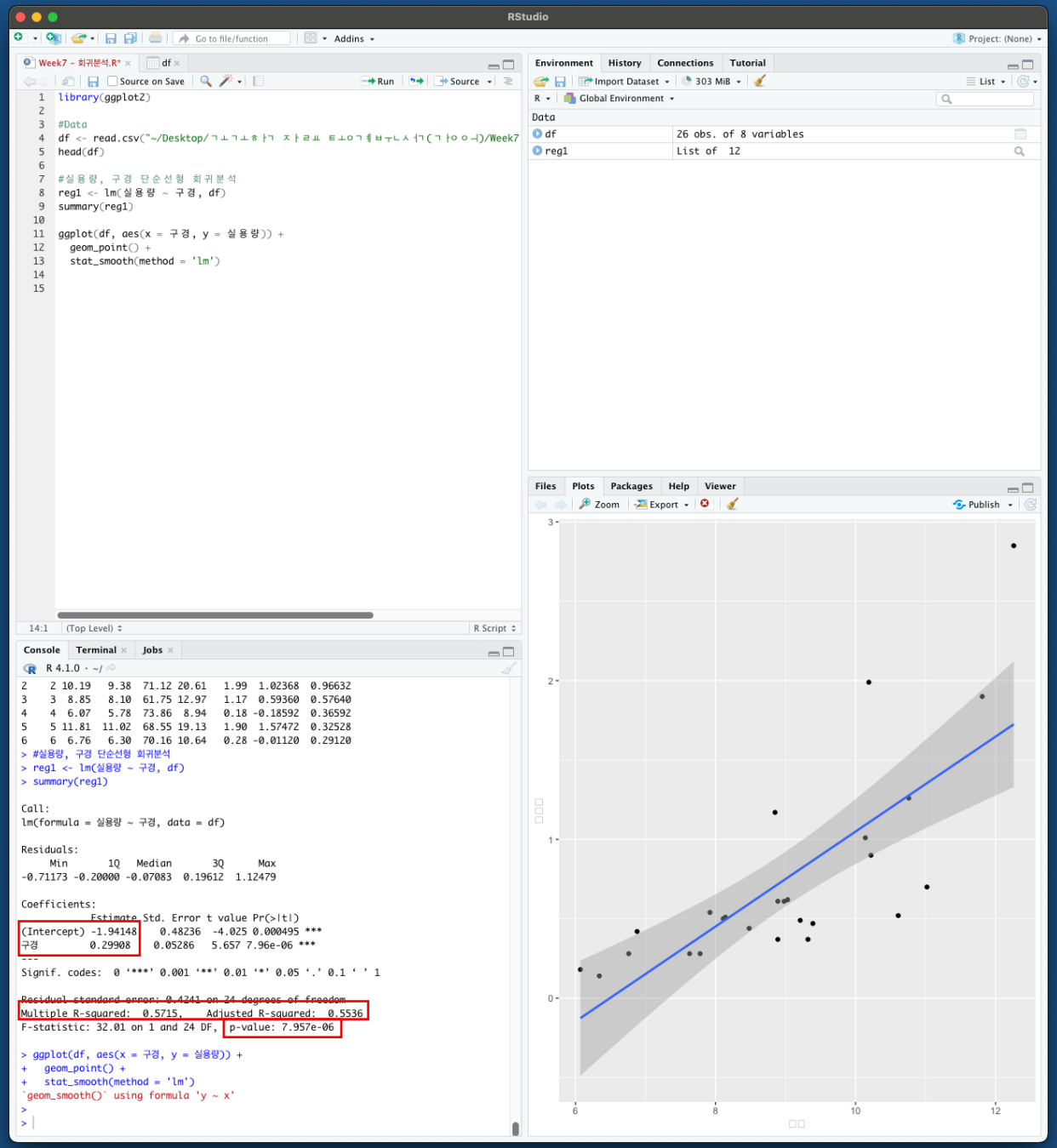
  Ex) $y = 0.29908x - 1.94148$

- 산출된 회귀계수를 통해 다른 자료 예측 가능

* R-squared : 설명력

* p-value : 유의성

* 이 경우엔 R-squared 값이 0.5715로 예측변수들의 57.15%밖에 설명하지 못함



단순선형 회귀분석

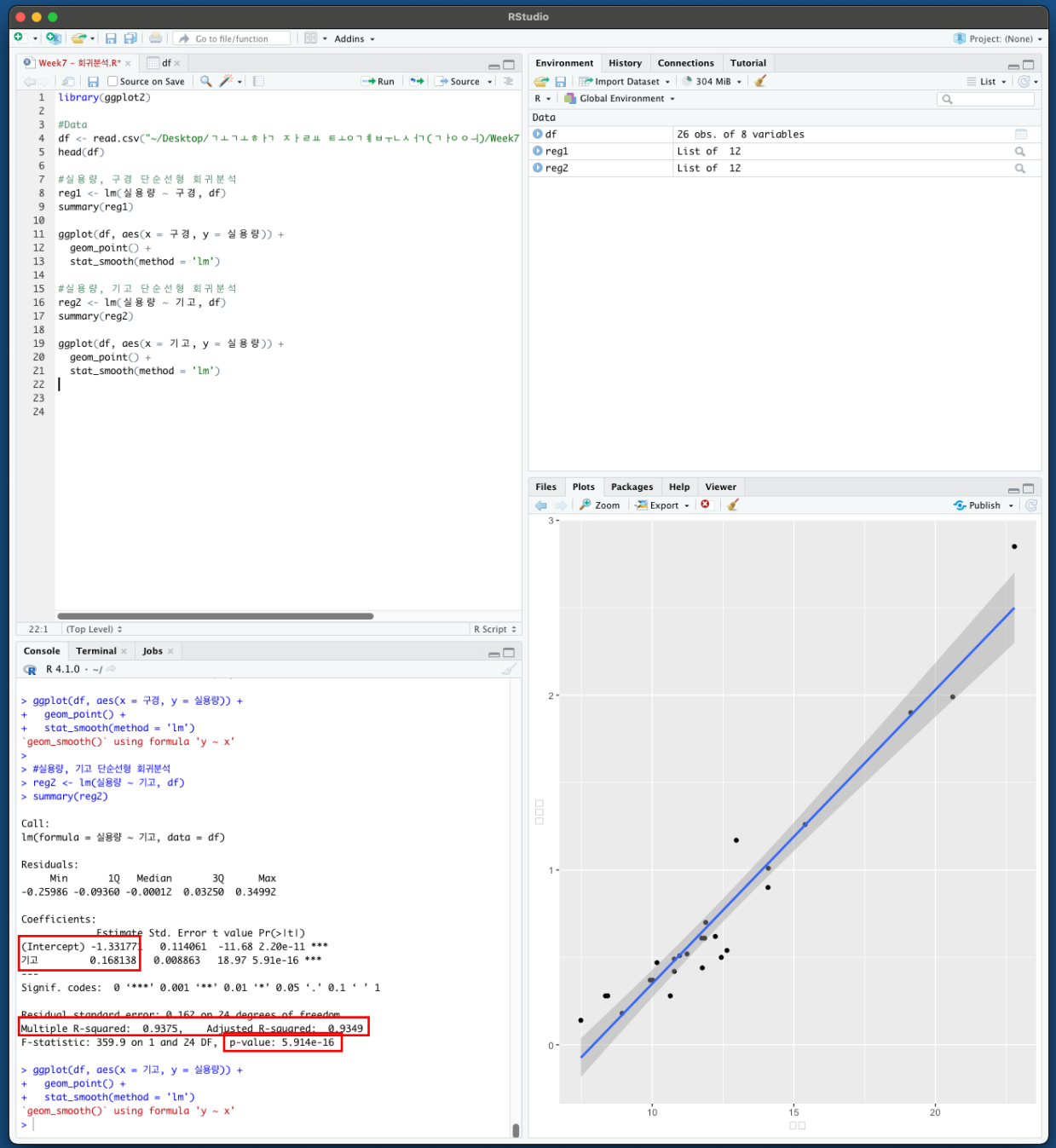
> 모델명 <- lm(y ~ x, 데이터)

> summary(모델명)

- lm() 명령어를 사용하여 회귀분석 실행
- summary를 통해 모델의 정보 출력
- 확인된 회귀계수와 상수를 통해 회귀식 산출 가능
- 산출된 회귀계수를 통해 다른 자료 예측 가능

* R-squared : 설명력
* p-value : 유의성

• 이 경우엔 R-squared 값이 0.9375으로 예측변수의 93%를 설명할 수 있음



다중선형 회귀분석

- 두 개 이상의 독립변수(x)가 하나의 종속변수(y)에 미치는 영향을 분석하는 방법
- **선형적 상관관계**를 모델링하여 분석하는 회귀분석의 일종.
- 장점
  - 여러 개의 변수를 활용할 수 있음
- 단점
  - 비선형적 변수는 제거해야함
  - 다차원일 경우 그래프로 표현할 수 없음

다중선형 회귀분석

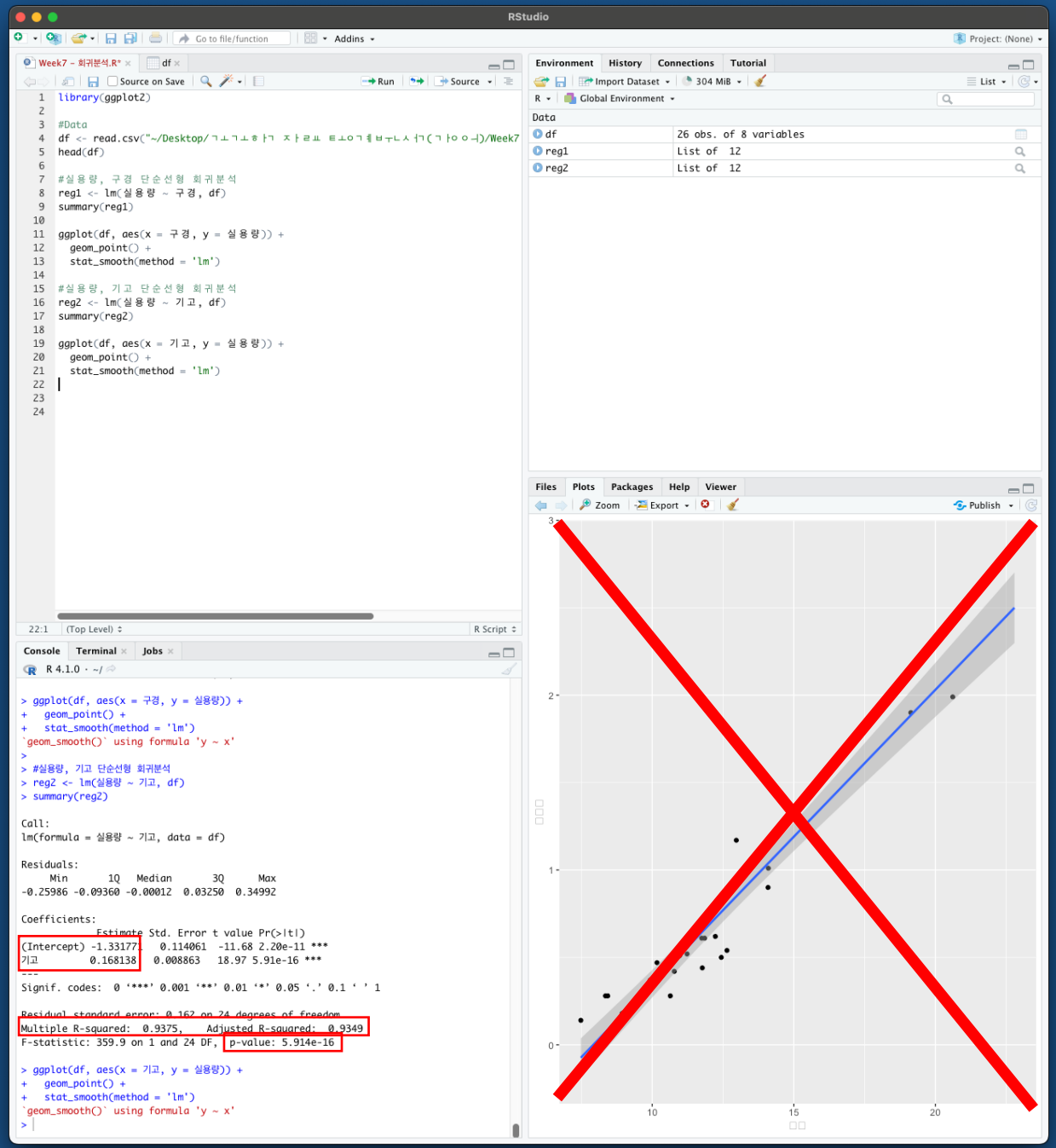
> 모델명 <- lm(y ~ x1 + x2 + x3 + … , 데이터)

> summary(모델명)

- lm() 명령어를 사용하여 회귀분석 실행
- 연구대상인 종속변수(y)는 그대로 두고 독립변수(x)에 변수들을 추가
- summary를 통해 모델의 정보 출력

* R-squared : 설명력
* p-value : 유의성

• 이 경우엔 R-squared 값이 0.9375으로 예측변수의 93%를 설명할 수 있음



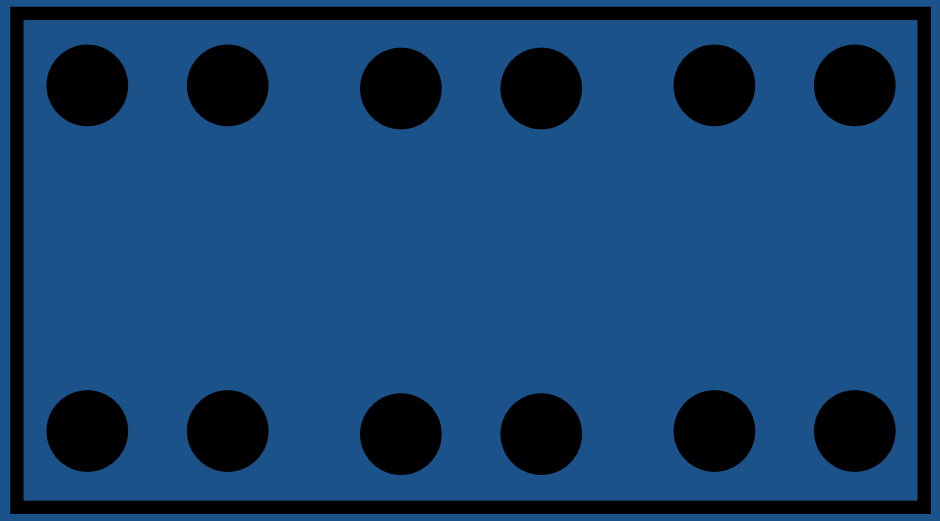
다중공선성(Multicollinearity)

- **독립변수 사이의 강한 상관관계**로 인해 분석 발생하는 부정적인 영향
- 회귀분석은 독립변수의 영향력이 일정하다고 가정하지만 두 독립변수가 서로에게 영향을 주고있을 경우 발생
- 다중공선성이 생기면 각각의 변수들의 **설명력이 약해짐**
- 분석결과의 유의성은 가설검정을 통해 판단(6주차) -> p-value가 유의수준보다 작아야함 -> p-value는 검정통계량이 클수록 작아짐 -> 설명력이 작아진 변수의 경우 p-value의 값이 커져 유의수준을 넘음
- 그러므로 **상관계수**와 **분산팽창요인**을 통해 다중공선성을 확인해야함
- 강한 상관관계를 보이는 경우 제거하여 해결

다중공선성(Multicollinearity)

*상황 가정
주거지 면적과 주구의 개수가 거주 인구에 미치는 영향에 대해 알고 싶음

독립변수 1 : 주거지의 면적
독립변수 2 : 주구의 개수

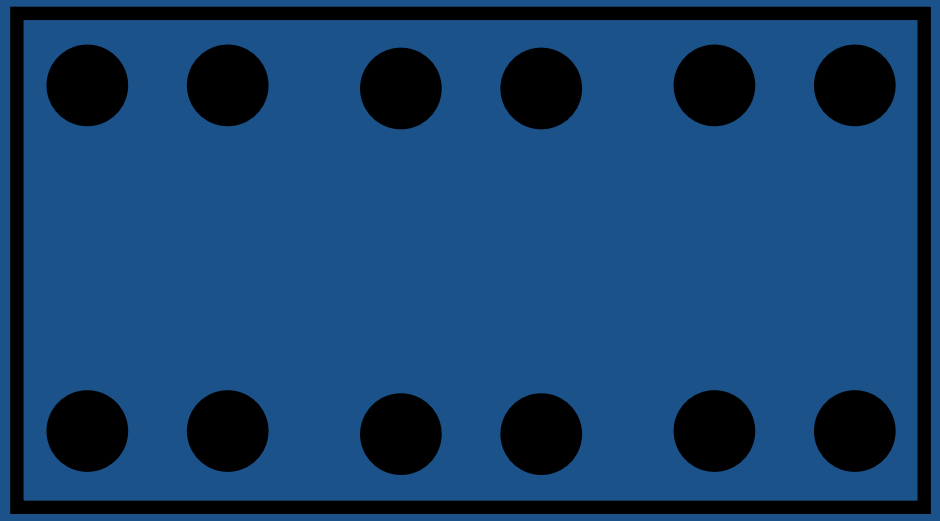


종속변수 : 거주 인구

다중공선성(Multicollinearity)

*상황 가정
주거지 면적과 주구의 개수가 거주 인구에 미치는 영향에 대해 알고 싶음

독립변수 1 : 주거지의 면적
독립변수 2 : 주구의 개수

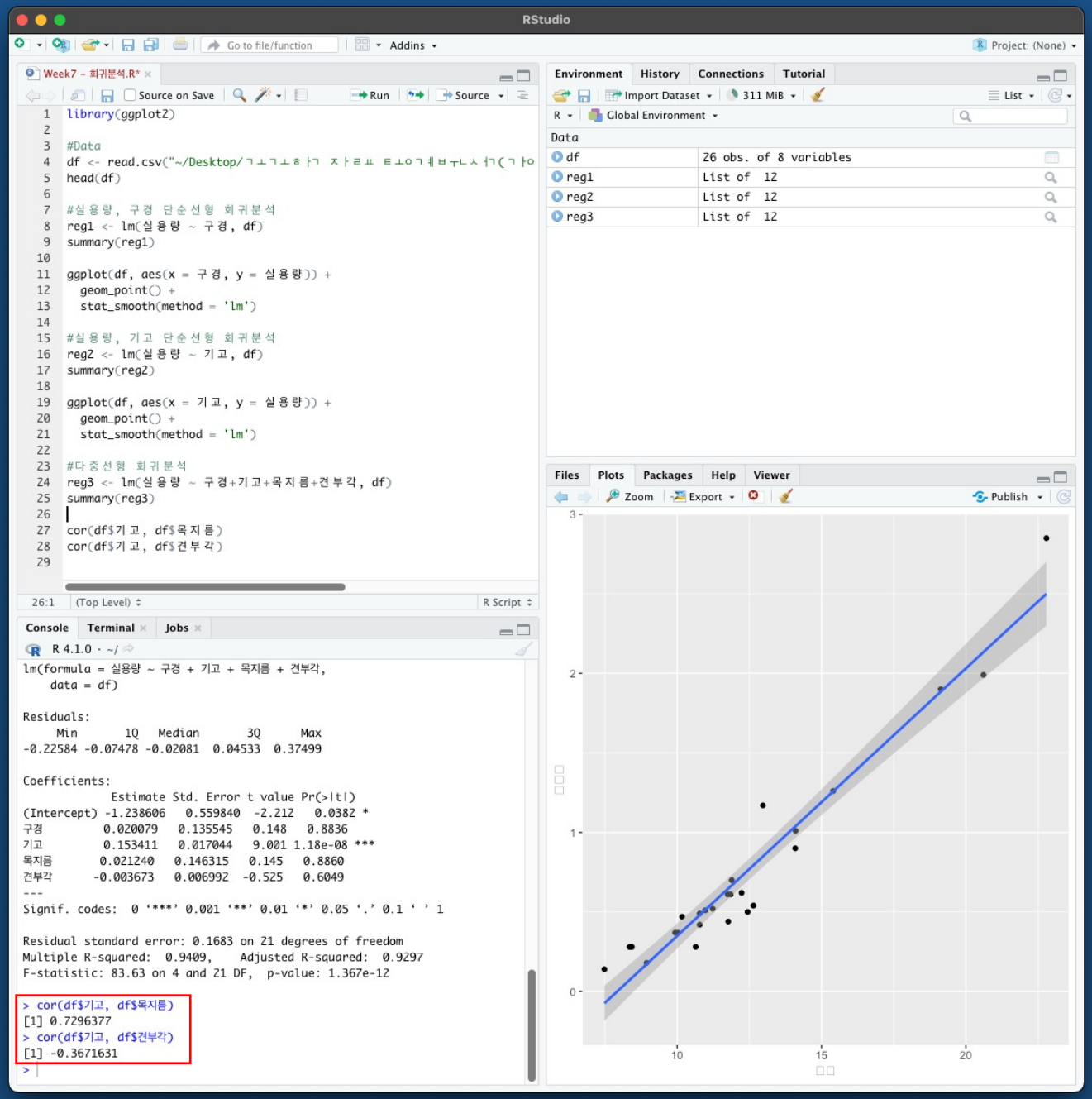


물리적으로 면적이 넓을수록
주구의 개수는 많을 수 밖에 없음!

상관계수

> cor(변수1, 변수2)

- cor() 함수를 사용하여 상관계수 산출
- 0.7 이상일 경우 다중공선성 존재

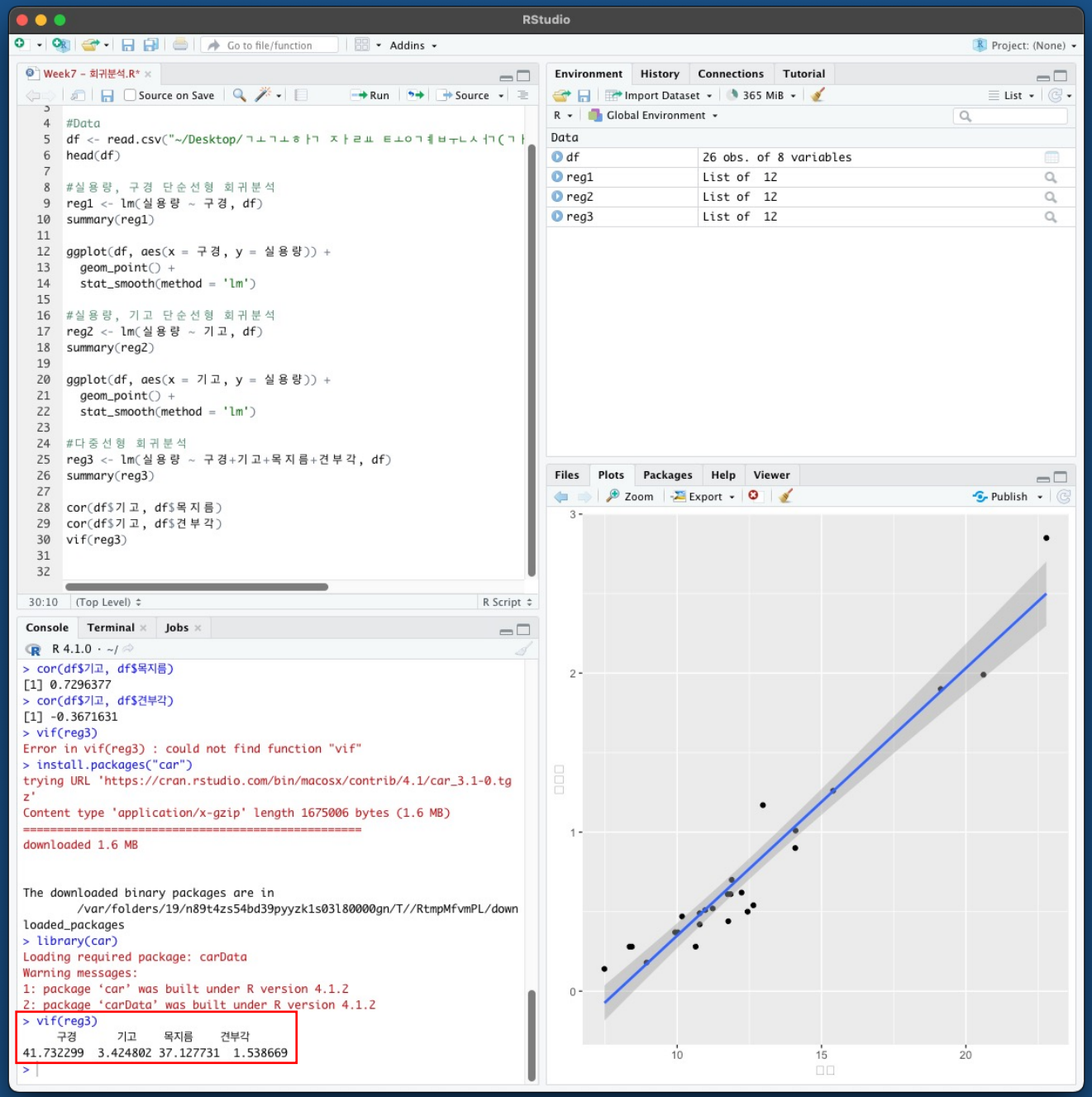


분산팽창요인

*한 독립변수가 다른 독립변수에 의해
설명되지 않는 부분을 역수로 표시한 것

```
> library(car)
> vif(회귀모델)
```

- vif() 함수를 사용하여 분산팽창요인 산출
- 10 이상일 경우 다중공선성 의심 필요



다중공선성 해결방법

- 상관계수나 분산팽창요인을 통해 다중공선성이 강하게 확인될 경우 해당 변수 제거
- 리지 회귀, 라쏘 회귀 등을 사용

로지스틱 회귀분석

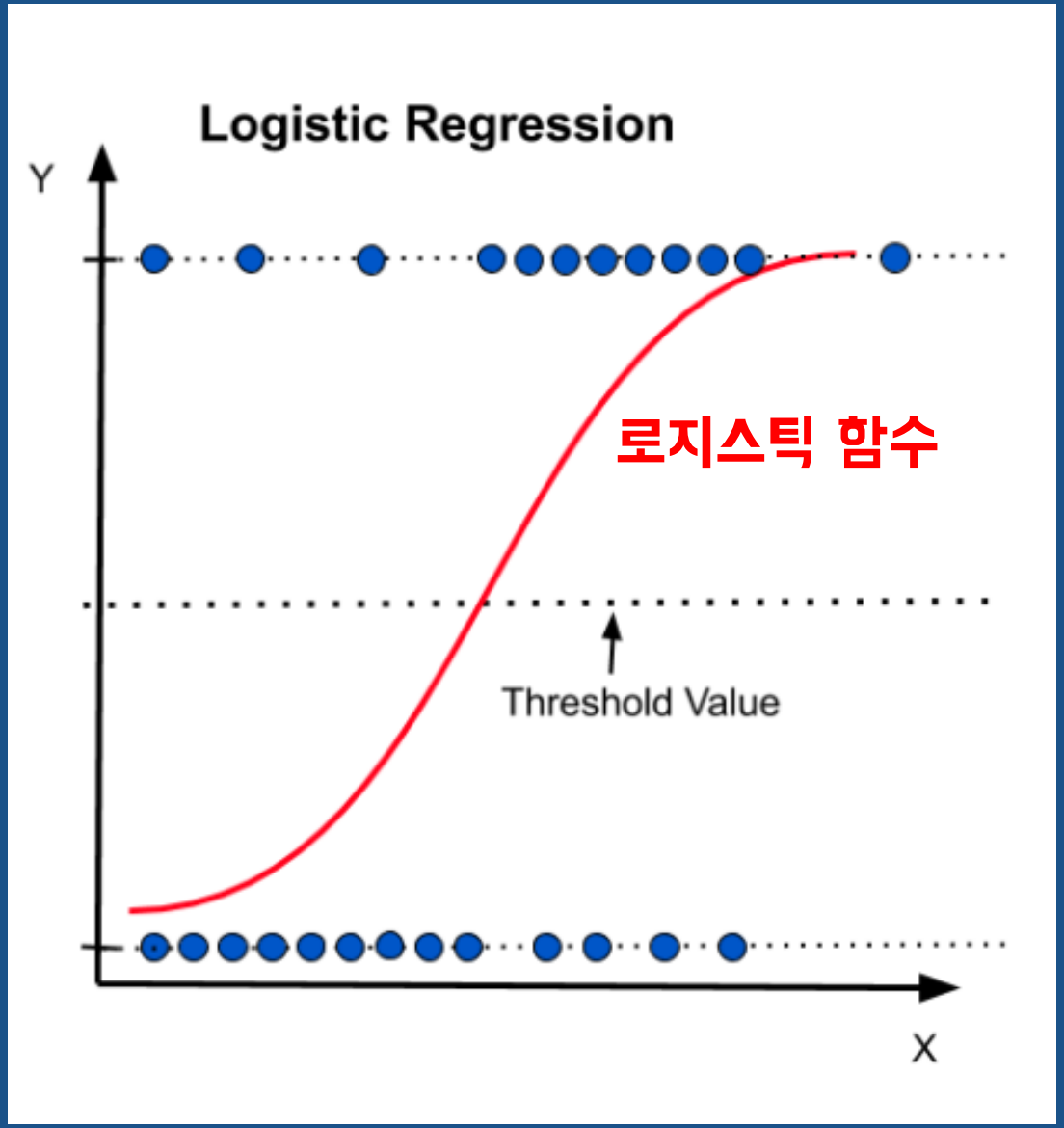
- **연속형** 변수인 독립변수(x)가 **범주형** 변수(이항형)인 종속변수(y)에 미치는 영향을 분석하는 방법
- 장점
  - 범주형 변수에도 사용 가능
  - 특성상 일종의 **분류(Classification)**로도 활용 가능
- 단점
  - 예측 성능이 다른 모델들에 비해 낮음
  - 해석이 어려움

로지스틱 회귀분석

- Odds : 한 사건이 어떠한 요인에 의해 발생하지 않을 확률 대비 발생할 확률(도박에서의 역배당 개념과 동일)
- Odds ratio : Odds 사이의 비율
- Logit : $\log(Odds\ ratio)$
- log를 사용하면 이항적인 종속변수를 음의 무한대부터 양의 무한대인 일반적인 연속변수로 바꿀 수 있음

로지스틱 함수

- 개체의 성장 등을 나타내는 함수
- 임계값(Threshold Value)를 기점으로 성장률은 둔화
- 대표적인 시그모이드 함수



로지스틱 회귀분석

> 모델명 <- glm(y ~ x, 데이터)

> summary(모델명)

- glm() 명령어를 사용하여 회귀분석 실행
- summary를 통해 모델의 정보 출력

* 이탈도(Deviance) :
Residual deviance Null deviance를 통해 변수의 영향력 판단

* ANOVA 검정을 통해서도 확인 가능