# 회귀분석팀


**6팀**

심은주
진수정
문병철
이수정
임주은

# INDEX

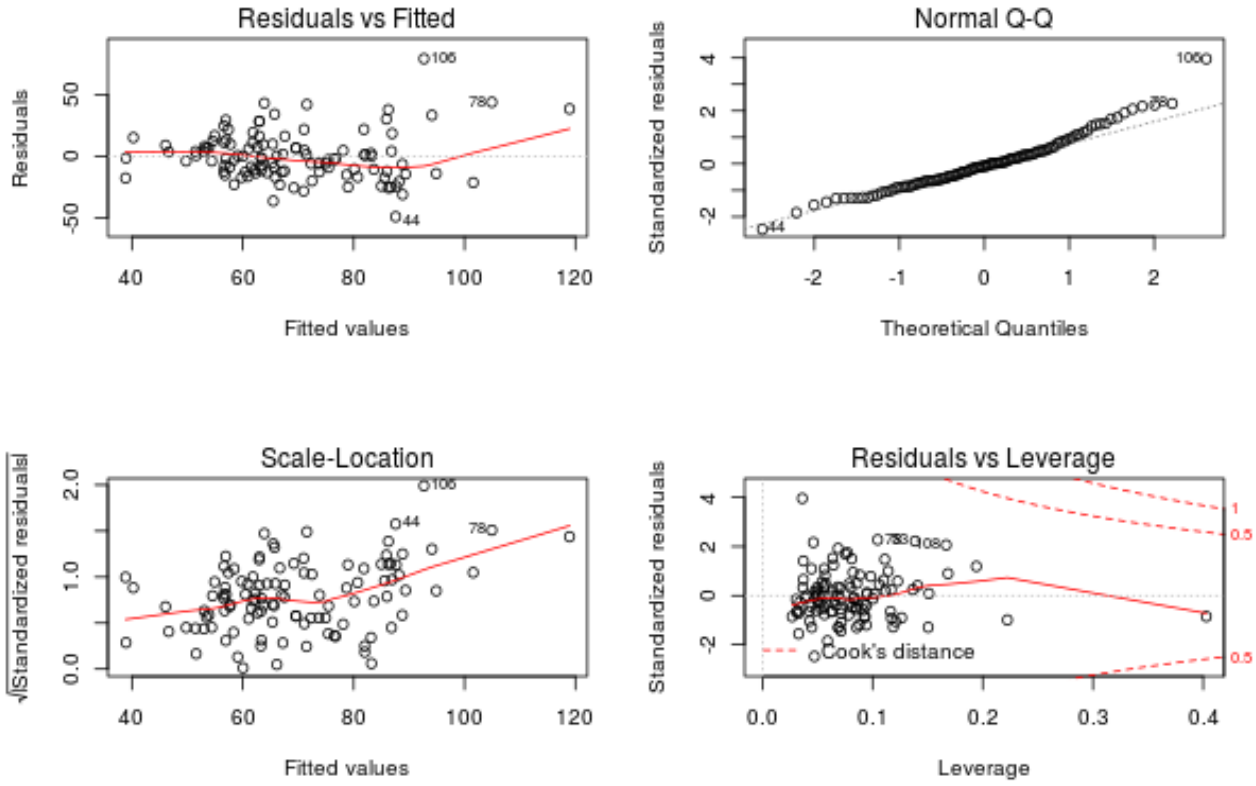
- 회귀분석의 가정

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \ , \ \varepsilon \sim NID(0, \sigma^2)$$

1. 식 자체가 X 변수들의 '선형결합'으로 이루어짐
2. 오차는 정규분포(N)를 따름
3. 오차들은 서로 독립적(ID)
4. 오차의 평균은 0, 분산은 $\sigma^2$

- **잔차플랏**

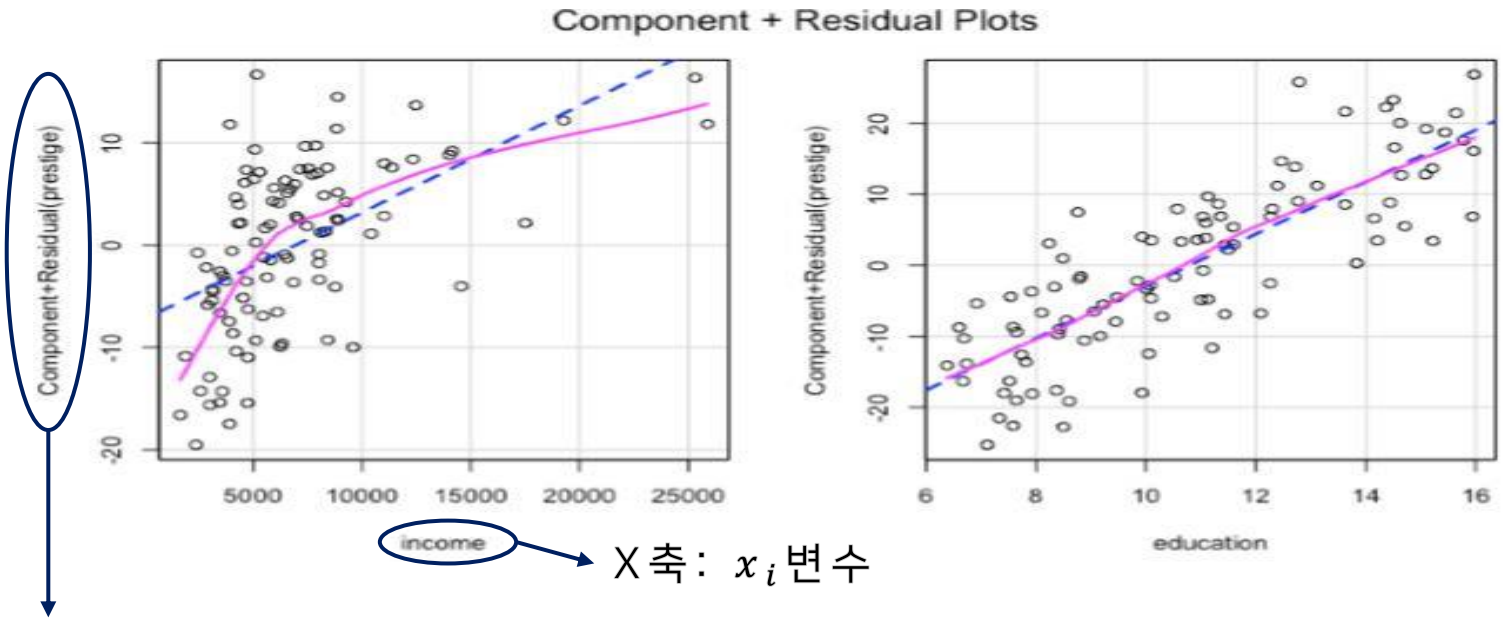
R에서 회귀식을 적합(fitting)하면 제공해주는 네 개의 플랏




회귀모형의 기본가정과 데이터의 문제를 시각적으로 진단 가능!

- 진단 - crPlots

Car 패키지의 CrPlots 함수를 통해 개별 변수의 선형성 파악




X 축 : $x_i$ 변수

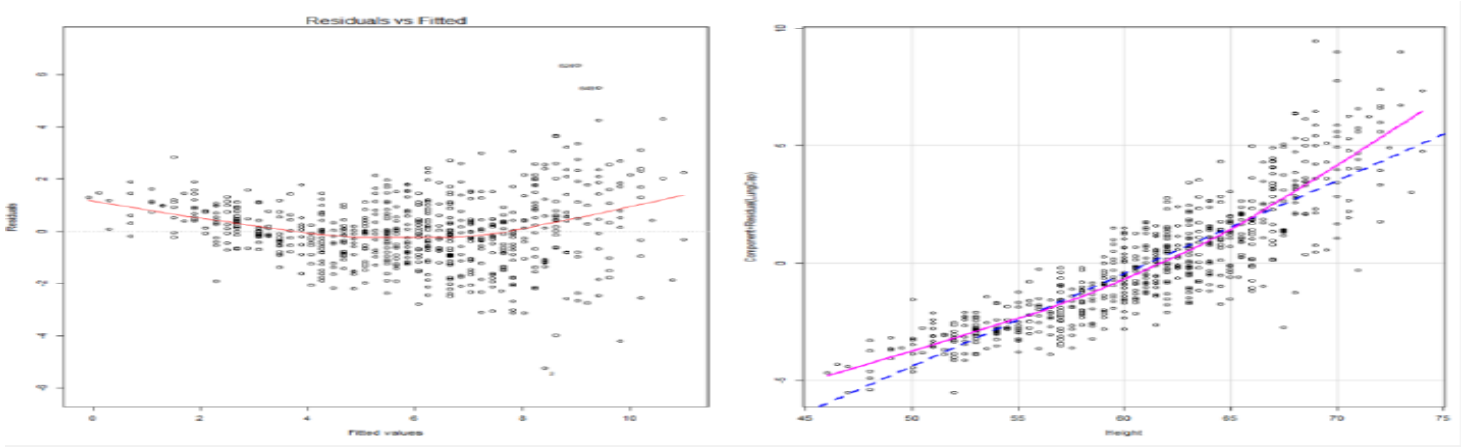
Y 축 : Partial residual ($y - (\beta_i x_i$를 제외한 모든 회귀식 성분))

파란 점선: Partial residual과 $x_i$의 적합된 직선

보라색 실선: 잔차의 추세선

- **처방 - Polynomial Regression**

잔차 플랏이나 Partial regression plot을 봤을 때, 2차 이상의 곡선 형태가 나타날 경우에 사용




Residual vs fitted plot
: 2차함수의 모양을 띔

분홍선(polynomial )이
데이터의 2차 모양을 잘 설명

**BUT** 3차를 넘어서는 모델링은 거의 하지 않음

- 진단 - test

**가설**

$H_0$: 주어진 데이터는 등분산성을 지님

$H_1$: 주어진 데이터는 등분산성을 지니지 않음

- BP(Breusch-Pagan) test : 가능도비검정 기반으로 추정

- 분산이 설명(X) 변수에 대한 선형결합으로 되어있음 가정

$$\sigma^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$$

- 위의 회귀식의 결정계수($R^2$) 값을 구해 검정통계량 계산

$$X^2_{stat} = nR^2 \sim X^2_{p-1}$$

- **처방 - 변수 변환**

  Y를 변환함으로써 등분산 혹은 정규성을 해결해주는 방법

  이때, 통계적인 검정에 따라 구한다는 점에서 효율적

Box-Cox Transformation

Yeo-Johnson Transformation

- 진단 - test

**가설**

$H_0$: 주어진 데이터는 정규분포를 따름

$H_1$: 주어진 데이터는 정규분포를 따르지 않음

### - Shapiro Wilk Test

R 기본함수로 내장되어 있으며, residual 값을 넣음

```
shapiro.test(salary.reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary.reg$residuals
## W = 0.96887, p-value = 1.41e-08
```

0.05기준으로 p-value가 작으니 귀무가설 기각
∴ 정규성을 만족하지 않음

- 진단 - test

***가설***

$H_0$: 잔차들이 서로 독립 (자기상관성이 없음)

$H_1$: 잔차들이 서로 독립이 아님 (자기상관성이 있음)

- Durbin Watson Test

바로 앞 뒤 관측치의 자기상관성을 확인하는 테스트

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \approx 2(1 - \widehat{\rho_1}), \qquad \widehat{\rho_1} = \frac{\sum_{t=2}^{n} e_i e_{i-1}}{\sum_{t=1}^{n} e_t^2}$$

$\widehat{\rho_1}$: 표본 잔차 상관, $e_i$와$e_{i-1}$의 상관계수의 꼴, $-1 \le \widehat{\rho_1} \le 1$

$$0 \le d \le 4$$

- **공간데이터의 특성: 공간자기상관**

공간상에 분포하는 객체 간의 상호작용

### *전역적 공간자기상관*
*Global Spatial Autocorrelation*

전체 구역이 가지는
하나의 공간자기상관의 정도

**예시** 한국 전체에서 나타나는
고혈압 유병률의 공간적인 패턴

### *국지적 공간자기상관*
*Global Spatial Autocorrelation*

개별 지점이 가지는
공간자기상관의 정도

**예시** 수도권에서 나타나는
고혈압 유병률의 공간적인 패턴

*공간 자기상관은…*

**특정 지역의 사건 강도가 인접 지역의 사건에 영향을 주는지를 파악하자!**

- **공간데이터의 특성: 공간적 이질성**

  넓은 지역에서 나타나는 불규칙한 분포를 의미하며,

  한 지역 내에 서로 다른 성격의 하위 집단이 존재하는 것을 말함

*Example*

지하철 개통이 지가에 미치는 영향력의 크기가 모든 지역에서 같은가?

→ 영향을 크게 받는 지역, 영향을 크게 받지 않는 지역 등 여러 유형의 집단 존재

*공간적 이질성은…*

**특정 사건이 전 지역에서 동일한 강도로 나타나는지 파악하자!**

- **공간자기상관 진단**

먼저 지역 내 지점들이 인접해 있는지부터 체크해야 한다!

공간가중행렬(Spatial Weights Matrix)

- 지역 내 다수의 지점들이 서로 공간적으로 인접하고 있는지의 여부를 파악할 수 있도록 행렬로 나타낸 것

- 지역 간의 잠재적 상호작용의 강도를 나타냄

$$w_{ij} = \begin{cases} 1 & \text{if } i,j \text{ is } neighbor \\ 0 & otherwise \end{cases}$$

이웃? 기준이 뭔데?

# 공간회귀모델 선택 알고리즘

모란 I 지수, LISA로 공간 자기상관성이 있는지 확인

라그랑지승수검정으로 모델 선택

*유의 X*

OLS 회귀모델

*LM-Lag 유의*

공간시차모델

*LM-Error 유의*

공간오차모델

*둘 다 유의*

Robust LM

*한 번 더 검정해서*

*더 유의한 모델 선택*

- **공간자기상관 처방**

| | |
|---|---|
| 공간 자기상관성 | 공간시차모델(SLM) |
| | 공간오차모델(SEM) |
| 공간적 이질성 | 지리가중회귀모형(GWR) |

**공간 자기상관성** → 인접지역의 영향력을 변수에 포함시켜 통제

**공간적 이질성** → 각 지역마다 다른 추정계수로 영향력을 추정