Regression

: X-Y 간 일차선형 관계 및 인과관계를 표현하는 식을 찾는 것

$$\hat{y} = \beta x + \beta_0$$

How to find? → 점추정

기본가정

① 오차들 e_i 는 확률변수이며 서로 독립 & 평균 0 분산은 σ^2

② Y는 서로 독립 & 평균 $\mu = \alpha + \beta X$, 분산 σ^2

$$L = \frac{1}{2}\Sigma(y_i - \hat{y}_i)^2 = \frac{1}{2}\Sigma(y_i - \beta x - \beta_0)^2$$

① $\frac{\partial L}{\partial \beta_0} = \Sigma(y_i - \beta_1 x - \beta_0)(-1) = 0$

$= -\Sigma y_i + \beta_1 \Sigma x + n\beta_0 = 0$

$\rightarrow \hat{\beta}_0 = \frac{1}{N}(\Sigma y_i - \beta_1 \Sigma x) = \bar{y} - \hat{\beta}_1 \bar{x}$

② $\frac{\partial L}{\partial \beta_1} = \Sigma(y_i - \beta_1 x - \beta_0)(-\Sigma x) = 0$

$= -\Sigma y_i \Sigma x_i + \Sigma \beta_1 x \Sigma x + \Sigma x \beta_0$

$= -\Sigma y_i \Sigma x_i + \Sigma \beta_1 x \Sigma x_i + \Sigma x_i (\bar{y} - \hat{\beta}_1 \bar{x})$

$= -\Sigma y_i \Sigma x_i + \hat{\beta}_1 \Sigma x_i \Sigma x_i + \bar{y} \Sigma x_i - \hat{\beta}_1 \bar{x} \Sigma x_i$

$\rightarrow \hat{\beta}_1 = \frac{\Sigma(y_i - \bar{y})}{\Sigma(x_i - \bar{x})} = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$

$\hat{\beta}_0, \hat{\beta}_1$ 로 식 구할 수 있다!

But 구한 식이 통계적으로 유의한지 검정이 필요!

How to find? → 구간추정

기본가정

① 오차들 e_i는 확률변수이며 $e_i \sim N(0, \sigma^2)$

② Y는 서로 독립 & $Y_i \sim N(\alpha+\beta X_i, \sigma^2)$

① β 추정

$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$ σ^2 모르므로 $S^2 = \frac{SSE}{n-2}$ 로 대체!

$$\left[\hat{\beta} - t_{\alpha/2, n-2}\sqrt{\frac{\frac{SSE}{n-2}}{S_{xx}}}, \hat{\beta} + t_{\alpha/2, n-2}\sqrt{\frac{\frac{SSE}{n-2}}{S_{xx}}}\right]$$

② α 추정

$\hat{\alpha} \sim N(\alpha, \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}))$

$$\left[\hat{\alpha} - t_{\alpha/2,n-2}\sqrt{\frac{SSE}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}, \hat{\alpha} + t_{\alpha/2,n-2}\sqrt{\frac{SSE}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right]$$

---

1) 위 식의 추정량은 "불편추정량"

α, β 추정해야되므로 $df = n-2$

σ^2의 추정량인 $S^2 = \frac{SSE}{n-2} = MSE$

$E(S^2) = E\left[\frac{SSE}{n-2}\right] = \sigma^2$ → bias X.

$S_{xx} = \Sigma(x_i - \bar{x})^2$ $S_{xy} = \Sigma(x-x_i)(y_i-\bar{y})^2$ $S_{yy} = \Sigma(y_i-\bar{y})^2$

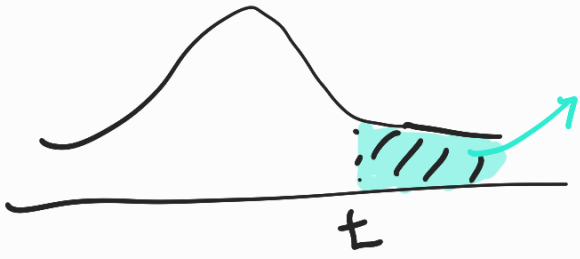
How to test?

① 변수 테스트 : $\hat{\beta}_1$에 대한 검정 (t-test)

$H_0 : \hat{\beta}_1 = 0$ → 변수의 설명력이 없다

$H_1 : \hat{\beta}_1 \neq 0$

$$\hat{\beta}_1 \sim N\left(0, \frac{s}{\sqrt{S_{xx}}}\right) \quad , \quad t = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{S_{xx}}}}$$

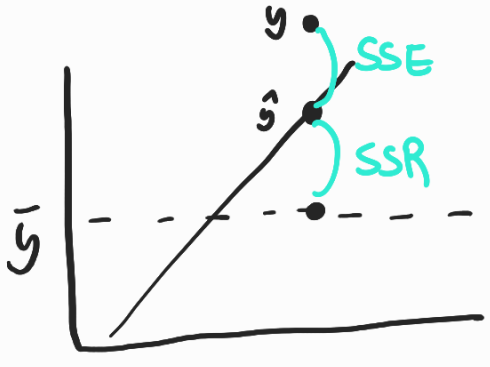


유의수준보다 작을 p-value 나올 경우 기각!

② 전체 회귀식 test : (F-test)

SSE : Sum of square Error
↳ 회귀식이 설명 못한 값

SSR : Sum of square regression.
↳ 회귀식이 설명하는 값



$H_0 : \beta_1 = \beta_2 = \cdots = 0$

$H_1 :$ 적어도 하나의 β는 0이 아님

$$F = \frac{MSR}{MSE} \sim \chi^2$$

	df	
SSR	p	MSR = SSR/p
SSE	n-p-1	MSE = SSE/n-p-1
TSS	n-1	

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \quad (0 \leq R^2 \leq 1)$$

R^2 0.6↑ 면 적합하다 판단

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

변수개수가 많아질수록 R^2값 올라가는데 이를 방지하기위해 R^2_{adj}

가정

① 선형성
: X-Y 관계가 선형적인지

② 독립성 (↔다중공선성)
: 다중회귀 모형에서 X변수들끼리 서로 연관 없는 것
: 오차항들은 서로 독립.

만족X? → 상관성 높은변수 제거 / 단계적 회귀분석 (전체 데이터를 몇개 구간으로 나눠 회귀분석)

* 자기상관 (auto correlation)
: 시계열 자료인 경우 시간에 대해 상관관계 발생
: 시계열 데이터의 독립성 판단

- 잔차 분포도 확인
  - 독립성 X면 t에 따라 패턴 존재
- 더빈왓슨 (DW) 통계량 ($0 \le DW \le 4$)
  - 2일 때가 상관계수 0이라 독립성만족
  - 0일 때 cor=1, 4일때 cor=-1

$$DW = \frac{\sum_{i=2}^{k} (e_i - e_{i-1})^2}{\sum_{i=1}^{k} e_i^2}$$

다중공선성 판정
① 모형 검정(F) ok (bad) 개별 변수 검정(t) 유의X
② 개별 변수 부호가 상식과 반대일 때
③ 상관분석 시 독립변수간 높은 상관성
④ VIF 계속

③ 등분산성

잔차의 분산이 일정 (독립변수 값에 상관없이 분산이 일정하다)

→ 오차항의 분산 σ^2이 균등해야한다

이분산성 판별

- 몇개 그룹으로 나누어 그룹간 분산 차이가 있는지 비교
- 잔차 그림 이용.

등분산성 X 등분산성 O

만족 X? → 변수변환

④ 정규성

: 잔차가 정규분포를 따른다

만족 X? → 변수 변환 (제곱근, x^2, 역수, 지수, 이차식..)