추정과 검정

추정과 검정

• 가설검정

- 모집단의 특성에 대한 가설을 설정하고 표본관찰을 통해 해당 가설을 채택할 지 여부를 결정하는 방법으로 언제나 귀무가설은 버릴 목적의 내용을 가설로 설정

(1) 귀무가설 (H0) : "비교대상의 값과 차이가 없다, ~와 같다"를 기본 개념으로 하는 가설

(거짓이 명확히 규명될 때까지 참인 것으로 인정되는 모수에 대한 주장)

(2) 대립가설(H1) : 귀무가설을 부정하는 가설, "~가 아니다" (귀무가설에 반대되는 가설)

보건복지부에서 전국의 100세 이상 노인의
평균 혈중 콜레스테롤이 174.6mg/dl라는 주장에 대한 검정

귀무가설 H0 : $m_0 = 174.6$ ⇨ 대립가설 H_1 : $\begin{cases} m_0 \neq 174.6 \\ m_0 > 174.6 \\ m_0 < 174.6 \end{cases}$

대립가설의 유형

- 대립가설의 유형

H0 : m = m_0 H_1 : m ≠ m_0	H0 : m ≤ m_0 H_1 : m > m_0	H0 : m ≥ m_0 H_1 : m < m_0
양측검정	상단측검정	하단측검정

[주의] 모수의 주장에 대한 = 부분은 항상 귀무가설에 포함시킨다.

가설검정_오류의 종류

- 검정결과를 모집단에 대한 것으로 일반화할 경우의 오류

실제상황 / 검정결과	H_0 가 참	H_1 이 참
H_0 채택	옳은 결정	제2종의 오류
H_1 채택	제1종의 오류	옳은 결정

- 제 1종 오류 α : 귀무가설 H0가 옳은데도 불구하고 대립가설을 채택하게 되는 오류
  (H0 를 기각하게 되는 오류)
- 제 2종 오류 β : 귀무가설 H0 가 옳지 않은데도 불구하고 H0 를 채택하는 오류

➢ 가설검정에서는 모든 오류가 작을수록 좋지만 모두 다 줄일 수 없는 관계에 있음

➢ 일반적으로 제 1종오류를 더 중요시 생각하고 제 1종 오류를 범할 확률의 최대 허용치를 미리 지정하고 검정

가설검정_유의수준

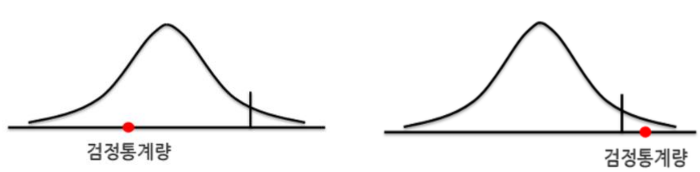
- 유의수준 (signification level)
  - 제 1종 오류를 범할 확률의 최대 허용치
  - 통상 α =0.05나 0.01 때때로 0.1까지 선택함. 즉 5%, 1%, 10%
  - α=0.05의 의미는 100번 검정 중 5번은 제 1종의 오류를 범한다는 의미

- 검정력(power of test)
  - 제 2종 오류(β)를 1에서 빼준 값
  - 1-β 는 틀린 귀무가설을 기각하여 귀무가설의 잘못을 찾아내는 확률
  - 검정력은 모수의 값에 따라 달라지는데, 이 함수의 값이 클수록 좋은 검정

- 검정통계량 (test statistic)
  - 관찰된 표본으로부터 구하는 통계량으로 검정 시 가설의 진위를 판단하는 기준임

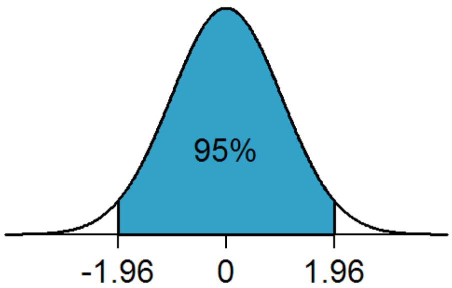
가설검정_기각역

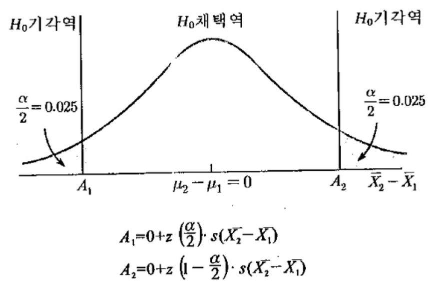
- 기각역 (Critical Region)
  - 검정통계량의 분포에서 유의수준 a의 크기에 해당하는 영역
  - 계산된 검정통계량의 유의성을 판정 하는 기준
  - 채택(accept) : H0이 타당하여 $H0$을 선택하는 경우, 귀무가설 $H0$을 채택한다 함
  - 기각(reject): 대립가설 $H1$이 타당하여 H0이 거짓인 경우, 귀무가설 $H0$을 기각한다 함
  - 임계값(critical value): 귀무가설을 기각시키거나 채택하는 범위를 구분하는 경계값



<귀무가설 채택> <귀무가설 탈락>

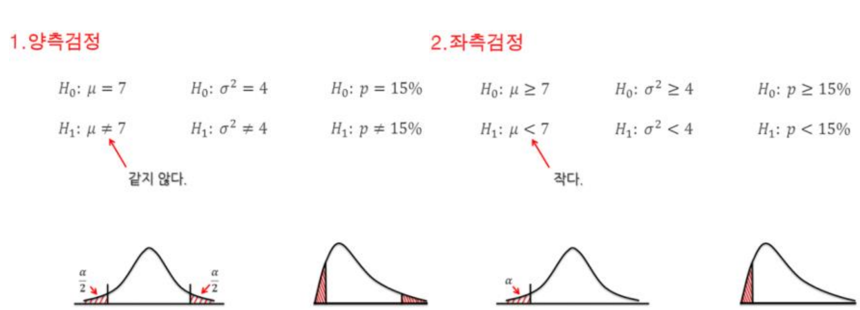
가설검정_기각역





가설검정_유의확률

- 유의확률(significance probability)
  - 관측치에 의해 귀무가설을 기각시킬 수 있는 검정법들의 유의수준 α 가운데 가장 작은 최소값 (흔히 유의확률을 p-value라 하고 유의수준 α보다 작으면 귀무가설을 기각하게 됨)
  - 양측검정인지 단측검정(우측 또는 좌측) 인지 유의



가설검정 절차

- 가설검정의 절차
  - 통계분석에서는 언제나 통계 값에 대해 p값을 확인하여 유의미한지 확인해야 함
  - 예를 들어 상관관계가 0.7이라고 하여 실제로 상관관계가 있다고 보는 것이 아니라
  - P값을 확인해보고 무엇이 문제인지 확인해봐야 하는 것

가설 검정의 절차

선형회귀분석
(Linear Regression)

회귀분석

- 회귀분석의 정의
  - 하나 혹은 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정하는 통계기법
  - 변수들의 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하여, 이 모형을 측정된 변수들의 데이터들로부터 추정하는 통계적 방법. 독립변수의 값에 의하여 종속변수의 값을 예측하기 위함

- 회귀분석의 변수
  - 영향을 받는 변수(Y) : 반응변수(response variable), 종속변수(dependent variable), 결과변수(outcome variable) 라고 함
  - 영향을 주는 변수(X) : 설명변수(explanatory variable), 독립변수(independentvariable), 예측변수(predictor variable)라고 함

회귀분석

- 단순 선형 회귀분석의 정의
  – 한 개의 종속변수와 한 개의 독립변수 간의 관계를 직선으로 표현하여 분석하는 방법

- 단순 선형 회귀분석의 모형

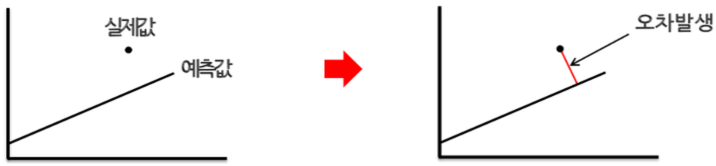
$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

β_0 : 절편 , β_1 : 기울기, ε : 잔차(residual)

  – 잔차 : 예측 값과 관측 값과의 차이
  – 회귀분석은 잔차를 최소화 할 수 있도록 절편과 기울기를 구함(최소자승법)
  – 회귀식 = 설명변수에 의해 설명되는 부분 + 설명되지 않는 부분 (오차)

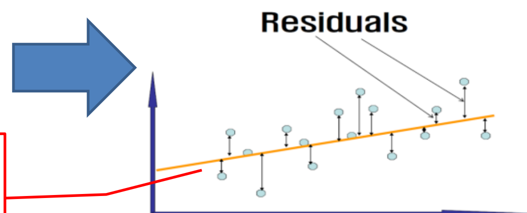
회귀분석

- 잔차 : 예측 값과 관측 값과의 차이



- 회귀분석은 잔차를 최소화 할 수 있도록 절편과 기울기를 구함(최소자승법)
  - 여러 잔차들이 최소가 되도록 하는 직선을 최소자승법을 이용해 찾음

$$e_i = y_i - \hat{y}_i$$



최소자승선은 데이터들 사이의 거리가
최소가 되도록 하는 직선식

회귀분석

- 잔차

$$e_i = y_i - \hat{y}_i$$

e_i : 잔차

y_i : (실제) 측정값

y_i hat : 예측값

- 오차

ε_i~i.i.d $N(0,\sigma^2)$

- 오차의 평균값은 0
- 정규성 : 오차 εi 는 정규 분포를 따름
- 독립성 : 오차 εi 는 서로 독립
- 등분산성 : 오차 εi 의 분산은 σ2 으로 동일

가우스 마코프 정리 : Gauss-Markov Theorem

- 오차변수의 기댓값은 0 이다.
- 오차변수와 독립변수의 공분산은 0이다.
- 오차변수의 분산은 일정한 상수이다.
- 오차변수들 사이의 공분산은 0이다.
- 오차변수는 정규분포를 따른다.<MVUE 가 되기 위한 조건>

- -> 이 조건을 만족할 때 최적 해

가우스 마코프 정리[참고 사이트]

- https://m.blog.naver.com/PostView.nhn?blogId=yunjh7024&logNo=220880125898&proxyReferer=https%3A%2F%2Fwww.google.com%2F

- https://m.blog.naver.com/PostView.nhn?blogId=yunjh7024&logNo=220881024021&targetKeyword=&targetRecommendationCode=1

단순선형 회귀분석 시 잔차에 대한 가정

- **단순선형 회귀분석을 사용하기 위해 잔차가 갖추어야할 네 가지 조건**
  - 오차의 평균값은 0
  - 정규성 : 오차 εi 는 정규 분포를 따름
  - 독립성 : 오차 εi 는 서로 독립
  - 등분산성 : 오차 εi 의 분산은 σ2 으로 동일

- **네 가지 조건을 만족해야 비로소 최소자승선은 예측치로 사용 가능**
- **네가지 조건을 만족했을 때 BLUE(Best Linear Unbiased Esitimator)라고 함**
- **네가지 조건 외에 오차변수가 정규분포를 따른다면 MVUE(Minimum Variance Unbiased Estimator)라고 부름**

→ **가우스 마코프 정리(Gauss-Markov Theorem)**

회귀분석

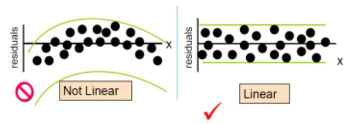
- 회귀계수의 추정
  - 최소자승법을 활용하여 잔차제곱 합을 최소로 하는 절편과 기울기를 추정
  - 적합된 회귀직선 : ŷ = b0+ b1x

- 회귀직선의 적합도 검토
  - 결정계수(R2)를 통해 추정된 회귀식이 얼마나 타당한지 검토
  - 독립변수가 종속변수 변동의 몇 %를 설명하는지 나타내는 지표
  - F 통계량

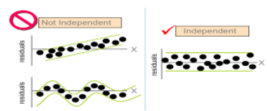
- 회귀분석의 장단점
  - 장점 : 결과를 통해 유효한 정보를 획득할 수 있고, 필요 없는 변수 선택을 통해 모델의 안전성을 높일 수 있음
  - 단점 : 사전에 결측치 처리 및 변수 간 교호작용의 유무 및 비선형 여부를 파악

회귀분석

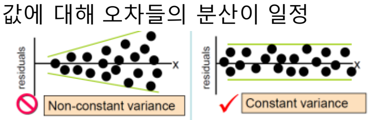
- 회귀분석의 가정
  - 선형성 : 독립변수의 변화에 따라 종속변수도 일정크기로 변화



  - 독립성 : 잔차와 독립변인의 값이 관련되어 있지 않음



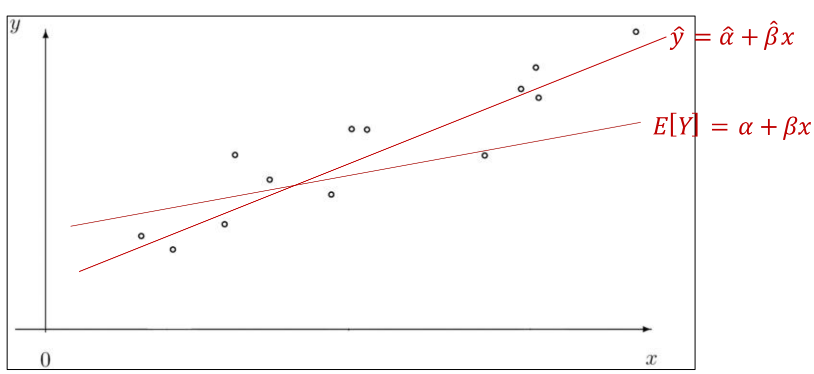
  - 등분산성 : 독립변인의 모든 값에 대해 오차들의 분산이 일정



  - 비상관성 : 관측치들의 잔차끼리 상관이 없어야 함
  - 정상성 : 잔차 항이 정규분포를 이루어야 함

단순 선형회귀

- 모수의 추정
  - 모형이 포함한 미지의 모수 $\alpha, \beta,\ \sigma^2$ 를 추정하기 위하여 각 독립변수 x_i 에 대응하는 종속변수 y_i 로 짝지어진 n 개의 표본 관찰치 $(x_i\ , y_i)$가 주어짐



단순 선형회귀

– 회귀계수 α와 β의 추정

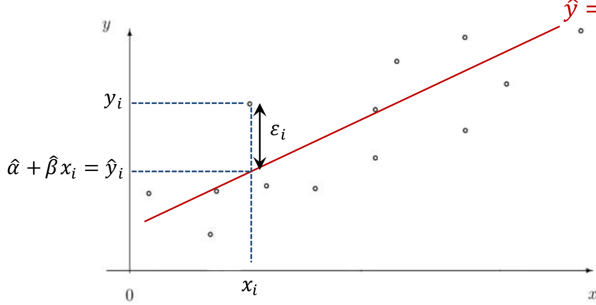
➤최소제곱법

단순회귀모형 $Y_i = \alpha + \beta x_i + \varepsilon_i$에서 오차의 제곱합

$$SS(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

이 최소가 되도록 α와 β를 추정하는 방법을 최소제곱법이라고 하고, 이 때 얻어지는 추정량 $\hat{\alpha}$, $\hat{\beta}$는 최소제곱 추정량이라고 함

$\hat{y} = \hat{\alpha} + \hat{\beta}x$ **: 추정된 회귀선**



최소자승선은 데이터들 사이의 거리가 최소가 되도록 하는 직선식

단순 선형회귀

➤ 최소제곱 추정량 $\hat{\alpha}$, $\hat{\beta}$ 의 도출

잔차의 제곱합(SS)이 최소가 되는 회귀계수 $\hat{\alpha}$, $\hat{\beta}$를 구하기 위해서는 SS를 α와 β로 편미분한 값을 0으로 둠

$$SS = \sum (y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial SS}{\partial \alpha} = \sum 2(y_i - \alpha - \beta x_i)(-1) = 0$$

$$\sum y_i = \sum \alpha + \beta \sum x_i$$

$$\sum y_i = n\alpha + \beta \sum x_i$$

$$\frac{\partial SS}{\partial \beta} = \sum 2(y_i - \alpha - \beta x_i)(-x_i) = 0$$

$$= \sum (y_i - \alpha - \beta x_i)(-x_i) = 0$$

$$= \sum x_i y_i = \alpha \sum x_i + b \sum x_i^2$$

단순 선형회귀

➢ 구해진 방정식을 정규방정식(Normal Equation)이라 하며, 회귀계수 추정량 $\hat{\alpha}$, $\hat{\beta}$는 이 정규방정식으로 구성된 연립방정식의 해로 도출할 수 있음

$$\hat{\alpha} = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} = \bar{y} - \beta\bar{x}$$

$$\hat{\beta} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

(단, $\bar{x}$는 x_i의 평균, $\bar{y}$는 y_i 의 평균)

➢ y_i의 추정치 : $\hat{y}_i = \hat{\alpha} - \hat{\beta}x_i \qquad (i = 1,2,\ldots,n)$

➢ 잔차 : $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i \qquad (i = 1,2,\ldots,n)$

단순 선형회귀

– 오차항의 분산 σ^2의 추정

➢ 오차에 대응되는 잔차의 변동성을 이용하여 추정

$SSE = \sum_{i=1}^{n} e_i^2$ 로 두고,

$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}$ 로 정의하면,

$E[MSE] = \sigma^2$ 임을 보일 수 있음

σ^2 의 추정량은 $\hat{\sigma}^2 = MSE$를 이용함

➢ SSE 의 간편 계산식

$$SSE = \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}\, x_i\right)^2 = S_{yy} - \hat{\beta}^2 S_{xx}$$

단순 선형회귀

- 추정량의 성질
  - 최소제곱 추정량 $\hat{\alpha}$, $\hat{\beta}$은 다음의 분포를 가짐

$$\hat{\beta} \sim Normal\left[\beta\ ,\quad \frac{\sigma^2}{S_{xx}}\right]$$

$$\hat{\alpha} \sim Normal\left[\alpha\ ,\quad \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right]$$

  - 오차항의 분산 추정량 $\hat{\sigma}^2 (= MSE)$은 다음의 분포를 가짐

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}\ x_i\right)^2}{\sigma^2} \sim \chi^2\ [\ n-2]$$

  - $\hat{\alpha}$, $\hat{\beta}$과 $\hat{\sigma}^2$은 서로 독립

단순 선형회귀

- 모형의 유의성 검정
  - 독립변수 x가 종속변수 Y를 설명하기 유용한 변수인가에 대한 통계적 추론은 회귀계수 β에 대한 검정을 통해 파악할 수 있음
  - t 검정
    - ➢가설

$$H_0\ :\ \beta = 0$$
$$H_1\ :\ \beta \neq 0$$

    - ➢검정통계량과 표본분포

귀무가설 H_0 이 사실일 때,

$$T = \frac{\hat{\beta}-0}{\widehat{S.E.[\hat{\beta}]}} \quad \sim \quad t\ [\ n-2], \qquad 단, \quad \widehat{S.E.[\hat{\beta}]} = \frac{\hat{\sigma}^2}{S_{xx}}$$

    - ➢기각역

$|T| = \left|\frac{\hat{\beta}-0}{\widehat{S.E.[\hat{\beta}]}}\right| \quad > t_{\alpha/2,n-2}$ 면 귀무가설을 기각

→ 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수라고 해석할 수 있음

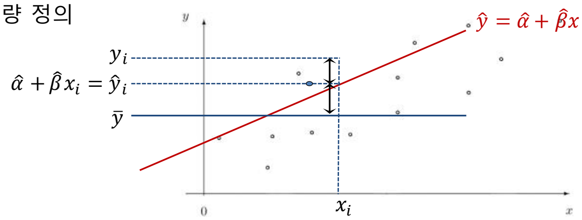
단순 선형회귀

– F 검정

➢가설

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

➢검정통계량 정의



y_i 의 변동을 추정된 회귀모형으로 설명되는 변동과 설명되지 않는 모형으로 분할

제곱합 :	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	+	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
	SST		SSR		SSE
	(y_i 의 변동)		(모형으로 설명되는 변동)		(모형으로 설명되지 않는 변동)
자유도 :	$(n-1)$	=	(1)	+	$(n-2)$

단순 선형회귀

– 귀무가설 H_0이 사실일 때, $MSR \approx MSE$이고, 대립가설 H_1이 사실일 때 $MSR \gg MSE$

→ 검정통계량을 $\frac{MSR}{MSE}$ 로 정의

➢검정통계량과 표본분포

귀무가설 H_0 이 사실일 때, $F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim F[1, n-2]$

➢기각역

$$F = \frac{MSR}{MSE} > F_{\alpha,1,n-2} \text{ 면 귀무가설을 기각}$$

➢분산분석표를 이용하여 결과를 정리

변동의 정의	SS 통계량	자유도	MS 통계량	검정통계량
회귀모형	SSR	1	MSR	F
오차	SSE	$n-2$	MSE	
전체	SST	$n-1$		

단순 선형회귀

- 모형의 적합성 검토
  - 결정계수 R^2
    - 결정계수 R^2 의 정의
    - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 로 정의
    - $SST = SSR + SSE$ 이므로 항상 0과 1 사이의 값을 가짐 $(0 \leq R^2 \leq 1)$.
    - y_i의 변동 가운데 추정된 회귀모형으로 통해 설명되는 변동의 비중을 의미함
    - 0에 가까울 수록 추정된 모형의 설명력이 떨어지는 것으로, 1에 가까울수록 추정된 모형이 y_i 의 변동을 완벽하게 설명하는 것으로 해석할 수 있음