presentation

robinhwp

2023-03-22

# 5강. 제2장 중회귀모형-2

## 1. 표준화된 중회귀분석

* 종속변수와 독립변수의 표준화 정규분포의 표준화변환 (통개학개론 p103)에서 $Z = { X - \mu \over \sigma }$ 식으로 변환하여 정규분포 을 따르게 되는 것처럼 변수의 표준화로 회귀계수()의 단위가 통일되어 비교가 가능해지게 된다.
  + 하지만 이 비교를 가지고 너무 단정적으로 결론을 내리는데는 주의하여야 한다.
  + $\begin{aligned} Y\_i^\* = {Y\_i - \bar{Y} \over \sqrt{S\_{YY}}} \end{aligned}$, $\begin{aligned} Z\_{ij} = {X\_{ij} - \bar{X\_j} \over \sqrt{S\_{jj}}} \end{aligned}$
    - : 번째 독립변수의 평균값
  + $\begin{aligned} \sum\_{i} Z\_{i\color{indianred}{j}} = 0, \sum\_{i} (Z\_{i\color{indianred}{j}})^2 = 1, (j=1,2,\cdots, k) \end{aligned}$

### 변수 표준화

* 표준화된 변수(standardized variables)의 중회귀모형
  + 표준화된 중회귀모형은 의 절편이 이 되는 특징이 있다
* 표준화된 중회귀모형 추정하면
  + 의 절대값이 크면 클수록 독립변수 가 에 주는 영향이 크게 된다.

## 2. 추정과 검정

### 추정된 회귀계수의 분산

* 회귀계수벡터 의 의 추정량   
  $\color{blue}{E(\boldsymbol{\hat{\beta}}) = \boldsymbol{\beta}}$
* 는 의 불편추정량
* 의 분산-공분산 행렬은  
  $\color{blue} {Var(\boldsymbol{\hat{\beta}})= (\boldsymbol{X'X})^{-1}\sigma^2}$

그런데 는 벡터 의 구성원들 간의 분산과 공분산을 나타내는 행렬로서 행렬의 구성원을  
라 하면 다음과 같이 쓸수 있다.

만약 우리가 특별히 관심 있는 독립변수 라면 의 분산이 작게되는 회귀방정식이 요구될 것이며,  
의 값이 크지 않도록 계획행렬(design matrix) 를 설계해 줄 필요가 있을 것이다.  
 행렬의 설계에 따라서 가 크게 달라지기 때문이다.

### 의 구간추정

* 독립변수들의 임의의 값 ()에서 의 구간추정

⇒

⇒ 의 % 신뢰구간  
$\color{blue}{\hat{Y} \pm t(n-k-1;\alpha/2)\sqrt{\boldsymbol{x'}(\boldsymbol{X'X})^{-1}\boldsymbol{x}\cdot MSE}}$

### 회귀계수 의 가설 및 검정통계량

* 가설
* 검정통계량
  + $\begin{aligned} t\_0={b\_i - \beta\_{i0} \over \sqrt{c\_{ii}\cdot MSE}} \end{aligned}$, 는 의 대각선 값

### 일반적 모형비교

* 두 모형의 비교는 완전모형과 축소모형의 잔차제곱합의 차이를 이용함 두 모형을 비교하기위한 검정통계량은 $\begin{aligned} F\_0 = {MSR \over MSE} \end{aligned}$ 을 사용한다.
* 완전모형 : 데이터에 잘 적합되리라고 고려되는 모형 ( )
  + 완전모형의 경우의 잔차제곱합
* 축소모형 : 귀무가설 의 가정하에서의 모형 ( )
  + 축소모형의 경우 잔차제곱합 ( 이라고 가정 )
  + 두 모형을 비교하기위한 검정통계량
    - $\begin{aligned} F\_0 = {MSR \over MSE} \end{aligned}$  
      $\begin{aligned} F\_0 = {{[SSE(R)-SSE(F)] \over df\_R - df\_F} \over {SSE(F) \\over df\_F}} &\Rightarrow {{[SST-SSE] \over (n-1) - (n-2)} \over {SSE \\over n-2}} \\&= {{SSR} \over {SSE \\over n-2}} = {MSR \over MSE} \end{aligned}$

## 3. 변수추가

### 추가제곱합

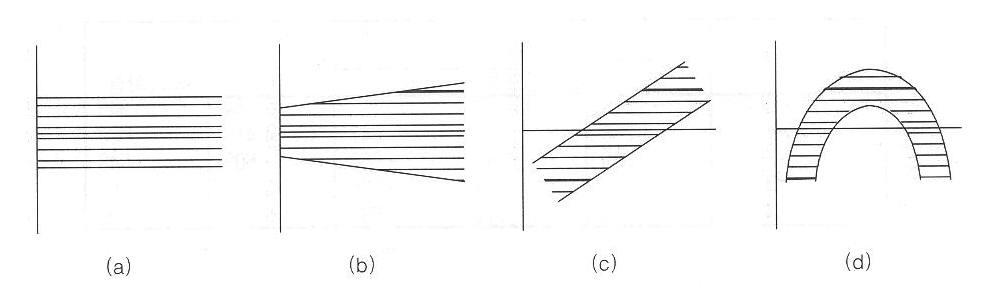
* 중회귀모형을 적합하는데 있어서 어떤 특정한 변수를 회귀모형에 포함시키는 것이 바람직한가를 결정하고 싶은 경우
  + 이 변수를 포함시키지 않고 구한 회귀제곱합보다 이 변수를 포함시키고 구한 회귀제곱합(regression sum of squares )이 추가적으로 어느 정도 커졌는가를 검토. 이와 같은 경우에 추가적으로 증가된 제곱합을 추가제곱합(extra sum of squares)이라고 함.
  + 추가제곱합은 새로운 변수가 모형에 추가될 때의 회귀제곱합의 증가분을 나타내는 것으로서 이 값이 작을수록 회귀에 대한 기여도가 떨어진다는 것을 의미.

### 추가변수그림(added variable plot)

* 중회귀모형에서 새로운 변수선택은 기존의 모형이 설명하지 못하는 부분을 새로운 변수가 들어옴으로써 추가설명력이 얼마나 유의한 가에 따라 결정
  + 새로운 변수의 효과를 그래프로 표현할 수 있는데, 이러한 그래프 중의 하나가 추가변수그림(added variable plot)임. 이를 편회귀그림(partial regression plot)이라고도 함.
* 추가변수 그림 그리는 절차
  + 독립변수가 2개인 회귀모형에서 변수 의 추가변수그림을 그리는 절차
  1. 를 으로 회귀한 후 얻어지는 잔차, 을 구한다.
  2. 를 으로 회귀한 후 얻어지는 잔차, 을 구한다.
  3. 앞에서 구한 두 잔차에서 -축을 , -축을 으로 한 산점도를 추가변수 그림이라 함.
* → 추가변수그림이 선형관계가 있으면 변수 는 추가적인 설명력이 있다고 판단.
* 위의 방식대로 강의에 있는 추가변수그림 첫번째 그래프를 그려보는 테스트 코드.
* # 실습코드 : 잔차 e(X1|X2,X3,X4)와 잔차 e(Y|X2,X3,X4)의 산점도  
  health = read.table("./data/health.txt", header=T)  
  # head(health,3)  
  h4.lm = lm(Y ~ X1+X2+X3+X4, data=health)  
  h4.lm.y1 = lm(Y ~ X2+X3+X4, data=health)  
  h4.lm.x1 = lm(X1 ~ X2+X3+X4, data=health)  
  h4.lm11=lm(resid(h4.lm.y1)~resid(h4.lm.x1))  
  plot(resid(h4.lm.x1),resid(h4.lm.y1), xlab="X1 | others", ylab = "Y | others")  
  abline(h4.lm11, col="red", lwd=2)

## 4. 잔차검토 및 분석 사례

### 잔차의 검토



**잔차(**)를 에 대한 산점도를 그렸을때,

ⓐ 가정에 아무런 모순이 없는것으로 판정된다.

ⓑ 분산이 일정하지 않으며, 가중회귀(weighted regression)를 쓰거나 또는 를 변환시켜 회귀분석함이 바람직하다.

ⓒ 절편이 필요한 모형인데 절편을 사용하지 않았을 경우에 생길 수 있는 형태이다.

ⓓ 모형이 타당하지 않다. 추가적으로 독립변수의 제곱항이 필요하다. 또는 의 적절한 변환이 필요하다.

**잔차(**)를 독립변수 에 대한 산점도를 그렸을때

ⓐ 가정에 아무런 모순이 없는 경우로 사용된 중회귀모형이 적절한다.

ⓑ 분산이 일정하지 않으며, 가중회귀(weighted regression)를 쓰거나 또는 를 변환시켜 회귀분석함이 바람직하다.

ⓒ 중회귀분석 과정에서 계산상에 착오가 있는 경우이다. 의 선형효과가 적절히 취급되지 않았다, 즉 항을 빠뜨려서 계산했거나 계산상에 실수가 있다.

ⓓ 필요한 독립변수의 항들이 포함되어 있지 않다. 특히 필요한 제곱항이 빠졌을 경우에 발생한다.