■ 공분산분석(Analysis of Covariance)

ex) 성별, 나이

- 요인(factor)과 공변량(covariate)을 설명변수로 갖는 선형모형

  ↳ concomitant variable

- ① 공변량 x가 반응변수를 예측하는데 어느 정도 기여할 수 있거나 확률화를 했으나 ② 각 처리에서 x가 균형을 맞추지 못한 경우

  ⇨ 모형에 공변량을 포함시킴으로써 실험의 정도와 검정력을 높을 수 있음

- ③ 확률화가 적용되지 않은 관측연구(observational study)의 경우 x가 골고루 섞여 있을 보장이 없음 ⇨ 모형에 공변량을 포함시킴으로써 편의(bias)를 줄임

  ① 직접 실험설계
  ② 관측만

○ 완전임의배치법

공변량 x가 pair로

요인			
1	2	…	k
(y_{11}, x_{11})	(y_{21}, x_{21})	$\cdots$	(y_{k1}, x_{k1})
(y_{12}, x_{12})	(y_{22}, x_{22})	$\cdots$	(y_{k2}, x_{k2})
$\vdots$	$\vdots$	$\ddots$	$\vdots$
(y_{1n_1}, x_{1n_1})	(y_{2n_2}, x_{2n_2})	$\cdots$	(y_{kn_k}, x_{kn_k})

$$Y_{ij} = \alpha + \overset{\text{처리효과}}{\tau_i} + \underline{\beta x_{ij}} + \overset{\text{모형의오차}}{\varepsilon_{ij}} = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij}$$

fixed effect model

- 가정: $\varepsilon_{ij} \sim$ iid $N(0,\sigma^2)$, $\sum \tau_i = 0$
- $\alpha = \mu - \beta \bar{x}_{..}$

복습) 변량효과모형에서는 모수가 고정된 값이 아니라 확률변수 (special case)
이런경우 아니면 모수는 모두 고정된 값이라고 생각!

$\sum_i\sum_j(\mu+\tau_i+\beta(x_{ij}-\bar{x}_{..}))$ 에서
- $\sum_i\sum_j \tau_i = 0$
- $\sum_i\sum_j (x_{ij}-\bar{x}_{..}) = \sum\sum x_{ij} - \bar{x}_{..}\times n = n\bar{x} - n\bar{x} = 0$

공변량때문에

- $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$

  모든 모수가 고정된 값, $\varepsilon_{ij} \sim N(0, \sigma^2)$

  - $E(Y_{ij}) = \mu_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{..})$ $\Rightarrow$ $\sum_i\sum_j \mu_{ij} = n\mu\,,\quad (n = n_1 + \cdots + n_k)$

    $= E(\mu+\tau_i+\beta(x_{ij}-\bar{x}_{..})+\varepsilon_{ij})$

    ↑ 모두 다른 반복수의 합

  - $E(\bar{Y}_{i.}) = \mu + \tau_i + \beta(\bar{x}_{i.} - \bar{x}_{..})\,,\quad E(\bar{Y}_{..}) = \mu$

    $= E\left(\frac{\sum_j Y_{ij}}{n_i}\right) = E\left(\frac{\sum_j(\mu+\tau_i+\beta(x_{ij}-\bar{x}_{..})+\varepsilon_{ij})}{n_i}\right)$

    $= E\left(\frac{\sum\sum Y_{ij}}{n}\right) = \frac{\sum\sum E(Y_{ij})}{n}$

- 공분산분석 모형에 대한 최소제곱추정량

  → 모형이 선형회귀모형이므로!
  수리통계의 MLE 처럼 회귀분석에서 사용
  내 data와 모델에서의 추정값의 차이 각각의 제곱합을 최소화하는 모수들의 추정치

  - $\hat{\mu} = \bar{Y}_{..}$
  - $\hat{\beta} = \dfrac{\sum_i\sum_j (x_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{Y}_{i.})}{\sum_i\sum_j (x_{ij} - \bar{x}_{i.})^2}$
  - $\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})$

  잔차들의 var

  - $\hat{\sigma}^2 = \dfrac{1}{n-k-1}\sum_i\sum_j\left(Y_{ij} - \bar{Y}_{i.} - \hat{\beta}(x_{ij} - \bar{x}_{i.})\right)^2$

$E(Y_{ij}) = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{..})$

$\rightarrow \hat{Y}_{ij} = \bar{Y}_{..} - \bar{Y}_{i.} - \bar{Y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) + \hat{\beta}(x_{ij} - \bar{x}_{..})$

$= \bar{Y}_{i.} + \hat{\beta}(x_{ij} - \bar{x}_{..})$

reduced model
full model

- 처리효과검정 $(H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0)$

차이가 없으면 reduced model 사용 (더 간단)

  - $SSE(R(\tau))$: 귀무가설(Reduced model) 하에서의 SSE
  - $SSE(F)$: 대립가설(Full model) 하에서의 SSE

reduce와 full의 차이가 지나치게 크면 정보손실 $\tau_1 \sim \tau_k$를 무시할 수 없음! → H_0기각 full model 선택

  - 검정통계량

reduced model의 자유도 $\begin{pmatrix} \alpha : 1 \\ \beta : 1 \end{pmatrix}$ → 2개추정필요 ⇒ $n-2$

$$F^* = \frac{SSE(R(\tau)) - SSE(F)}{(n-2)-(n-k-1)} \Big/ \frac{SSE(F)}{n-k-1} \sim F_{k-1,n-k-1}$$

full model의 자유도 $(Y_{ij} = \alpha + \tau_i + \beta x_{ij} + \epsilon_{ij})$ $\begin{pmatrix} \alpha : 1 \\ \beta : 1 \\ \tau_i : k-1 \end{pmatrix}$ → $k+1$개의 parameter 수 추정 ⇒ $n-(k+1) = n-k-1$

- 기울기 β에 대한 검정 $(H_0 : \beta = 0)$ → 공변량 x가 영향을 미치지 않음

  - 검정통계량

$\begin{pmatrix} \alpha : 1 \\ \beta : 0 \\ \tau_i : k-1 \end{pmatrix}$ → k개 ⇒ $n-k$

$$F^* = \frac{SSE(R(\beta)) - SSE(F)}{(n-k)-(n-k-1)} \Big/ \frac{SSE(F)}{n-k-1} \sim F_{1,n-k-1}$$

- μ_{ij}에 대한 추정: $E(Y_{ij}) = \mu_{ij} \Leftarrow \hat{Y}_{ij} = \hat{\alpha} + \hat{\tau}_i + \hat{\beta} x_{ij}$

공변량이 처리에 따라 다른 영향을 줄 수 있다는 점을 고려

- 기울기의 동일성(test of parallel slopes)

$$Y_{ij} = \alpha + \tau_i + \beta_i x_{ij} + \varepsilon_{ij}$$

㊂

  - 기울기의 동일성: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$
  - $Y_{ij} = \alpha + \tau_i + \beta + (\tau\beta)_i + \varepsilon_{ij}$ 에서 상호작용 $(\tau\beta)_i$ 에 대한 유의성 검정
  - Full 모형과 Reduced 모형의 SSE 비교

· reduced model
$\alpha(1), \beta(1), \tau_i(k-1) \rightarrow k+1$
· full model
$\alpha(1), \beta(k), \tau_i(k-1) \rightarrow 2k$

$$F^* = \frac{SSE(R) - SSE(F)}{\underbrace{k-1}_{(n-k-1)-(n-2k)}} \Big/ \frac{SSE(F)}{n-2k} \sim F_{k-1, n-2k}$$

  - 기울기의 동일성 가정이 성립 ⇨ 공분산분석 (H0 채택)
  - 기울기의 동일성 가정 할 수 없음 ⇨ 일반선형모형을 통해 각 처리수준에서의 기울기 추정하고 기울기에 대한 차이 및 전반적 현상을 해석 (H0 기각) → 복잡!

cf) 차원축소
정보의 손실없이 독립성을 유지하며

11/22