■ 공분산분석(Analysis of Covariance)

- 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 → x의 차이에서 오는 잘못된 결론

  ⇨ 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

공변량 x가 페어로 measure됨

○ 완전임의배치법

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})

선형

$$Y_{ij} = \alpha + \tau_i + \beta x_{ij} + \varepsilon_{ij} = \mu + \underset{\uparrow\,\text{처리효과}}{\tau_i} + \boxed{\beta(x_{ij} - \bar{x}_{..})} + \varepsilon_{ij}$$

fixed effect model로 돌아옴

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

$Var(Y_{ij}) = Var(\mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij})$
$= Var(\varepsilon_{ij}) = \sigma^2$

공변량에서 중요한 점

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

- $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$
  - 모두 상수
  - $E(Y_{ij}) = \mu_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}) \Rightarrow \sum_i \sum_j \mu_{ij} = n\mu, \quad (n = n_1 + \cdots + n_k)$
    $= E(\mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij})$
  - $E(\bar{Y}_{i.}) = \mu + \tau_i + \beta(\bar{x}_{i.} - \bar{x}_{..}), \quad E(\bar{Y}_{..}) = \mu = E\left(\frac{\sum\sum Y_{ij}}{n}\right) = \frac{\sum\sum E(Y_{ij})}{n}$
    $= \left(\sum_j (\mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}))\right) / n_i$
    $= \frac{\sum\sum(\mu + \tau_i + \beta(x_{ij} - \bar{x}_{..}))}{n}$
- 공분산분석 모형에 대한 최소제곱추정량
  - $\hat{\mu} = \bar{Y}_{..}$
  - $\hat{\beta} = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{Y}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}$
  - $\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})$

regression 형태에서 많이 사용
내 data 와 모델에서의 추정값의 각각 차이를 제곱함
이것을 최소화하는 모수들의 추정치

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

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

차이가 없으면 reduced model로 가야함 (더 간단)

- 처리효과검정 $(H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0)$
  - $SSE(R(\tau))$: 귀무가설(Reduced model) 하에서의 SSE
  - $SSE(F)$: 대립가설(Full model) 하에서의 SSE
  - 검정통계량

full model에서는
$\alpha \to 1$
$\tau_i \to k-1$
$\beta \to 1$
→ k+1개
의 파라미터를 추정해야함 → 자유도는 $n-(k+1) = n-k-1$

크면 H_0 기각 → full model 선택

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

reduced model에서는 $\tau_i = 0$이므로 날아감
$\alpha \to 1$
$\beta \to 1$
→ 자유도는 $n-2$

- 기울기 β에 대한 검정 $(H_0 : \beta = 0)$: 영향을 안미침
  - 검정통계량

reduced model 에서는 $\beta = 0$
$\alpha \to 1$
$\tau_i \to k-1$
→ k개, 자유도는 $n-k$

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

full model에서는 $\beta \neq 0$
$\alpha \to 1$
$\tau_i \to k-1$
$\beta \to 1$
→ k+1개, 자유도는 $n-k-1$

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

$= \hat{\mu} + \hat{\tau}_i + \hat{\beta}(x_{ij} - \bar{x}_{..})$

- 기울기의 동일성(test of parallel slopes) 좀더 general한 모형

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

  - 기울기의 동일성: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$

    가설검정

    $Y_{ij} = \alpha + \tau_i + \beta + (\tau\beta)_i + \varepsilon_{ij}$ 에서 상호작용 $(\tau\beta)_i$ 에 대한 유의성 검정과 같다.
  - Full 모형과 Reduced 모형의 SSE 비교

SSE(R): $\alpha, \tau_{1\sim k-1}, \beta$

↳ $k+1$개, 자유도 $n-k-1$

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

SSE(F): $\alpha, \tau_{1\sim k-1}, \beta_{1\sim k}$ → $2k$개, 자유도 $n-2k$

  - 기울기의 동일성 가정이 성립 ⇨ 공분산분석 (H_0 채택)
  - 기울기의 동일성 가정 할 수 없음 ⇨ 일반선형모형을 통해 각 처리수준에서의 기울기 추정하고 기울기에 대한 차이 및 전반적 현상을 해석 (H_0 기각)