Efekty losowe

Bogumiła Koprowska Elżbieta Kukla

Jednokierunkowa ANOVA

Było: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

$i = 1, 2, \ldots, k$,
$j = 1, 2, \ldots, n$
Y_{ij} - j-ta obserwacja na i-tym poziomie czynnika
μ - ogólna wartość średnia
α_i - efekt i-tego poziomu czynnika
ε_{ij} - niezależne zmienne losowe $\sim N(0, \sigma^2)$

Obserwacje Y_{ij}, $j = 1, 2, \ldots, n$ dla każdej ustalonej wartości wskaźnika i tworzą **grupę**.

Dane zgrupowane

- każda obserwacja należy do jednej grupy i jest tylko jeden czynnik grupujący
- hierarchiczne, zagnieżdżone zbiory danych - bardziej skomplikowane przypadki
- nie możemy zakładać niezależności obserwacji, obserwacje są skorelowane w ramach grup
- efekty losowe - wygodne do modelowania tego typu danych

Efekt losowy a efekt stały

1. **Efekt stały**
   nieznana liczba, którą estymujemy z danych

2. **Efekt losowy**
   zmienna losowa, estymujemy parametry opisujące rozkład efektu losowego

Przykłady

1. **Z astronomii:**
   1861 r. - astronom Airy sformułował model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ z losowymi efektami α_i do opisu obserwowanych przez teleskop obiektów astronomicznych

2. **Z medycyny:**
   traktowanie pacjentów poddanych leczeniu, jako wybranych losowo z większego zbioru pacjentów, których cechy chcemy estymować

Kiedy używać efektów losowych?

- gdy efektom danego czynnika nie można przypisać charakteru stałego, należy uznać je za losowe
- gdy chcemy uzyskać informacje o całej populacji, nasze obserwacje traktujemy jako próbkę z całej populacji
- niektórzy statystycy polecają zawsze używanie efektów losowych

Model mieszany

- w modelu mieszanym pojawiają się efekty stałe i efekty losowe
- najprostszy przykład - **model dwukierunkowej klasyfikacji**:

$$y_{ijk} = \mu + \tau_i + \nu_j + \varepsilon_{ijk}$$

gdzie

$i = 1, 2, \ldots, m, \ j = 1, 2, \ldots, l, \ k = 1, \ldots, n_{ij}$

efekty stałe:
$\mu, \ \tau_i$

efekty losowe, i.i.d.:
$\nu_j \sim N(0, \sigma_\nu^2)$
$\varepsilon_{ijk} \sim N(0, \sigma^2)$

Estymacja efektów, przykład

1. **Hipoteza zerowa** przy estymacji efektów:
   - stałych - $H_0 : \tau_i = 0 \; \forall i$
   - losowych - $H_0 : \sigma_\nu^2 = 0$
2. **Przykład - badania biologiczne:**
   badanie cechy potomków zadanej populacji I matek, np.
   efekt czynnika płci - efekt stały,
   efekt j-tej matki - efekt losowy

Estymacja

Estymacja efektów

Estymacja efektów w modelu losowym jest możliwa przy użyciu:

- klasyfikacji jednokierunkowej (ANOVY)
- metody największej wiarogodności (ML)
- metody największej wiarogodności z restrykcjami (REML)

Najprostszy model losowy

Model jednokierunkowej klasyfikacji (ANOVA):

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$i = 1, \ldots, a$
$j = 1, \ldots, n$
$\alpha_i \sim N(0, \sigma^2_\alpha)$
$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$

- **Komponenty wariancyjne**: σ^2_α, σ^2_ε
- **Wewnątrzgrupowy współczynnik korelacji**: $\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon}$

Estymatory

- **Dekompozycja wariancji:**
  $\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2$

  $SST = SSE + SSA$

- $E(SSE) = a(n-1)\sigma_\varepsilon^2$
  $E(SSA) = (a-1)(n\sigma_\alpha^2 + \sigma_\varepsilon^2)$

- **Estymatory:**
  $\hat{\sigma}_\varepsilon^2 = SSE/(a(n-1)) = MSE$
  $\hat{\sigma}_\alpha^2 = \frac{SSA/(a-1)-\hat{\sigma}_\varepsilon^2}{n} = \frac{MSA-MSE}{n}$

Wady ANOVY

Wady korzystania z ANOVY:

- estymator wariancji może przyjmować ujemne wartości
- skomplikowane obliczenia w przypadku gdy dane są niezbalansowane

Metoda największej wiarogodności (ML)

- nie ma takich wad jak ANOVA
- musimy założyć, jaki rozkład mają błędy i efekty losowe (zwykle rozkład normalny)
- dla modelu wyłącznie z efektami stałymi mamy:

$$y = X\beta + \varepsilon,\ y \sim N(X\beta, \sigma^2 I)$$

gdzie
X - macierz planu doświadczenia *nxp*
β - wektor p stałych parametrów
ε - wektor błędów losowych

Model mieszany

- dla modelu mieszanego:

$$y = X\beta + Z\gamma + \varepsilon,\ y|\gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$

gdzie
Z - macierz nxq opisująca wpływ efektów losowych na obserwacje y
γ - wektor q efektów losowych
ε - wektor błędów losowych

- dalej zakładamy, że $\gamma \sim N(0, \sigma^2 D)$, wówczas:

$$vary = varZ\gamma + var\varepsilon = \sigma^2 ZDZ^T + \sigma^2 I$$
$$y \sim N(X\beta, \sigma^2(I + ZDZ^T))$$

- przyjmuje się, że wektory losowe γ i ε są niezależne

Estymacja przy użyciu metody największej wiarogodności (ML)

- jeśli przyjmiemy, że $V = I + ZDZ^T$, wówczas łączna gęstość y wynosi:
$$\frac{1}{(2\pi)^{n/2}|\sigma^2 V|^{1/2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T V^{-1}(y-X\beta)}$$
- logarytm wiarogodności:
$l(\beta, \sigma, D|y) = -\frac{n}{2} log 2\pi - \frac{1}{2} log|\sigma^2 V| - \frac{1}{2\sigma^2}(y - X\beta)^T V^{-1}(y - X\beta)$
- estymacji podlega nieznany wektor stałych parametrów modelu β, wariancja σ^2 i macierz komponentów wariancyjnych D

Wady ML

Wady korzystania z metody największej wiarogodności (ML):

- trudności w przypadku bardziej złożonych modeli, wymagających estymacji większej liczby paremetrów efektów losowych
- problem z estymatorem wariancji, gdy maksimum wiarogodności jest przyjmowane dla ujemnych wartości
- estymatory są często dość silnie obciążone

Metoda największej wiarogodności z restrykcjami (REML)

- nie ma wielu wad, które dotyczyły ANOVY i ML
- dla danych zbalasowanych, estymatory obliczone przy użyciu REML są zwykle takie same jak obliczone przy użyciu ANOVY
- dostajemy estymatory nieobciążone lub o zredukowanym obciążeniu
- idea: bierzemy K - dopełnienie ortogonalne X (tzn. $K^T X = 0$).

  Wówczas: $K^T y \sim N(0, \sigma^2 K^T V K)$,

  następnie możemy korzystać z metody ML (nie mamy już efektów stałych)

Testowanie

Test ilorazu wiarogodności

Do porównywania hipotez H_0 i H_1 używamy testu **ilorazu wiarogodności**:

$$2(l(\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1|y) - l(\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0|y))$$

gdzie

$\hat{\beta}_1, \hat{\sigma}_1, \hat{D}_1$ - estymatory największej wiarogodności (MLE) parametrów przy hipotezie alternatywnej

$\hat{\beta}_0, \hat{\sigma}_0, \hat{D}_0$ - MLE przy hipotezie zerowej

Wady testu ilorazu wiarogodności

Wady testu ilorazu wiarogodności:

- test jest przybliżony i często przybliżenie to jest nienajlepsze
- wymaga wielu dodatkowych założeń, np. parametry przy hipotezie zerowej nie mogą być na brzegu przestrzeni parametrów

Testowanie efektów stałych

- nie możemy użyć metody REML do porównania dwóch modeli, które różnią się tylko efektami stałymi
- do testowania efektów stałych używamy ML (jeśli chcemy korzystać z testu ilorazu wiarogodności)
- p-wartości generowane przez test ilorazu wiarogodności dla efektów stałych są przybliżone i często zbyt małe - wyolbrzymiają wagę niektórych efektów
- do znalezienia dokładniejszej p-wartości możemy użyć metody **parametryczego bootstrapu**

Testowanie efektów losowych

- hipoteza zerowa zwykle ma postać $H_0 : \sigma^2 = 0$
  problem - to nie jest wnętrze przestrzeni parametrów
- jeśli jednak przyjmiemy, że otrzymujemy rozkład χ^2, wówczas otrzymana p-wartość będzie zwykle większa niż powinna
- metody numeryczne (bootstrap)

Dziękujemy!