Eksploracja Danych

Analiza Wariancji (ANOVA)

(c) Marcin Sydow

Zawartość wykładu

- Analiza wariancji
- Analiza wieloczynnikowa
- Testy post-hoc
- Kontrasty
- Analiza wieloczynnikowa
- Interakcje

Ogólny model

Analiza modelu o ogólnej postaci:

- $$y|X \sim \mathcal{F}(\Theta)$$
- $$E(y|X) = f(X, \beta)$$

y - zmienna objaśniana (atrybut decyzyjny)
X - zmienne objaśniające (atrybuty)

Obserwacji podlegają wartości zmiennej losowej o rozkładzie z rodziny $\mathcal{F}(\Theta)$ indeksowanej parametrem Θ

Przykład: $\mathcal{F}$ to rodzina rozkładów normalnych o parametrach o wariancji σ i wartości oczekiwanej $E(y|X)$

Zadanie: opisanie wartości oczekiwanej zmiennej y za pomocą atrybutów X, tj. wyznaczenie parametrów β przy założonym modelu f.

W zależności od typu atrybutów można dokonać następującej kategoryzacji procedur statystycznych:

- Metody analizy wariancji (ang. ANalysis Of VARiance - ANOVA): atrybut decyzyjny: ilościowy, atrybuty (zmienne objaśniające): kategoryczne. Ocena czy średnie wartości zmiennej y istotnie różnią się pomiędzy grupami obserwacji wyznaczonymi przez różne wartości atrybutów kategorycznych
- Metody analizy regresji
- zmienne objaśniające zarówno kategoryczne jak i ilościowe: ANCOVA (ANalysis of COVAriance) an. war. ze zmiennymi towarzyszącymi)

Liczba zmiennych objaśniających:

- jednoczynnikowa(jednokierunkowa): jedna zmienna
- dwukierunkowa: dwie zmienne
- wielokierunkowa: wiele zmiennych

Atrybut decyzyjny jednowymiarowy:

- jednowymiarowy: ANOVA
- wielowymiarowy: MANOVA (Multi-variate ANOVA), wielowymiarowa analiza wariancji, rodzaj kombinacji ANOVA i analizy regresji

Podstawy analizy wariancji

Obserwacji podlega n obiektów (pomiarów), które można pogrupować ze względu na wartość pewnego (lub pewnych) atrybutów kategorycznych (nominalnych).

Celem jest ustalenie, czy wartość średnia atrybutu (ilościowego) y różni się w poszczególnych grupach.

Podstawowe założenia

Zakładamy, że $E(y|X = a) = \mu + \mu_a$, gdzie:

- μ to średnia **bazowa** atrybutu decyzyjnego y
- μ_a to efekt dla grupy obiektów o wartości atrybutu grupującego wynoszącej a.
- dodatkowo, zakłada się (normalizacja): $\sum_a \mu_a = 0$

Typowe dodatkowe założenia: wewnątrz danej grupy y ma rozkład normalny o wariancji σ^2.

Przykłady zastosowań

- skuteczność leczenia w zależności od rodzaju terapii
- wysokość zarobków w zależności od specjalizacji
- cena mieszkania w zależności od dzielnicy
- wydajność plonów w zależności od zastosowanego rodzaju nawozu

Analiza jednoczynnikowa

Zależność pomiędzy ilościowym atrybutem decyzyjnym a jendą zmienną jakościową.

Zadanie: sprawdzić, czy wartość średnia atrybutu y różni się istotnie w zależności od wartości zmiennej nominalnej posiadającej k poziomów.

W języku testowania hipotez:
$H_0 : \mu_1 = \mu_2 ... = \mu_k$

μ_i to wartość średnia atrybytu decyzyjnego y w grupie i

$H_A : \exists_{i,j} \mu_i \neq \mu_j$

Założenia

- zgodność szumu z rozkładem normalnym
- niezależność szumu od wartości atrybutów

Należy sprawdzić powyższe założenia (stosując metody m.in. testowania zgodności, etc.)

Przykład

Dane dotyczące cen mieszkań w pewnym mieście.

Atrybut decyzyjny y to cena mieszkania.

Załóżmy, że występują m.in następujące atrybuty kategoryczne:

- dzielnica miasta
- typ budynku
- liczba pokoi

flats = read.table("flats.csv")

summary(flats)

Przykład c.d.

Do tego zadania wykorzystać można metody analizy wariancji np. do uprzednio dopasowanego modelu liniowego:

linModD = lm(cena~dzielnica, data= flats)

summary(linModD)

linModT = lm(cena~typ.budynku, data = flats)

summary(linModT)

im niższa p-wartość tym silniejsza przesłanka za odrzuceniem hipotezy zerowej

Przykład c.d. obiekt R anova

Obiekt typu anova jest to ramka danych posiadająca następujące atrybuty:

- $Df - liczba stopni swobody
- $"Sum Sq" - suma kwadratów wartości wyjaśnionych przez daną zmienną (używana do obliczenia statystyki F)
- $"Mean Sq" - średnia suma kwadratów
- $"F value" - wartość statystyki testowej F
- $"Pr(>F)" - p-wartość testu F

Przykład

anova(linModD)

anova(linModT)

Na poziomie istotności co najmniej $\alpha = 0.01$ można uznać, że ceny w poszczególnych kategoriach się różnią.

Testy post-hoc

Po odkryciu, że wartości się różnią w analizie wariancji, można przystąpić do kolejnych testów, które pokazują, **które średnie się różnią** i jak.

Służą do tego tzw. testy post-hoc, porównujące różnice parami, m.in.:

- test HSD Tukeya
- test Scheffé'a
- test Dunnetta
- test Newmana-Keulsa
- test Ryana
- test Duncana
- test Fishera
- test WSD Tukeya.

Przykład: test HSD Tukeya


```
a1 = aov(cena~dzielnica, data = flats)
a2 = aov(cena~typ.budynku, data = flats)
TukeyHSD(a1)
TukeyHSD(a2)
plot(TukeyHSD(a1))
plot(TukeyHSD(a2))
plot(cena~dzielnica,data = flats)
plot(cena~typ.budynku,data = flats)
```

Kontrasty

Ogólne testy post-hoc porównują wszystkie pary grup.

Do porównywania wybranych grup służą tzw *kontrasty*.

Kontrast to liniowa funkcja średnich μ_i:

$$L = \sum_{i=1}^{k} c_i \mu_i$$

taka, że suma współczynników wynosi zero: $\sum_{i=1}^{k} c_i = 0$

Jeśli średnie są sobie równe (hipoteza zerowa) to wartość kontrastu wynosi zero.

Dobór odpowiedniego kontrastu (tj. wartości czynników c_i) pozwala na porównywanie wybranych grup.

Przykład kontrastu

Można też definiować własne kontrasty, np:

$$L = -\mu_1 + 2\mu_2 - \mu_3$$

w R, np:

contr = cbind(c(-1,2,-1),c(-1,-1,2)), etc.

pozwala na porównanie wartości μ_2 z pozostałymi wartościami średnimi.

Funkcje do Tworzenia Grup Kontrastów w R

Kontrasty w grupie powinny być ortogonalne (tzn. reprezentujące je wektory współczynników c, k: liczba czynników).

Przykłady predefiniowanych kontrastów w R:

- contr.treatment: pierwsza średnia traktowana jako bazowa, wzystkie pozostałe są z nią porównywane, $L_i = \mu_i$ (R: contr.treatment(k), domyślny w)
- contr.sum: porównania do ostatniego czynnika, $L_i = \mu_i - \mu_k$ (R: contr.sum(k))
- contr.helmert: porównanie średniej z i pierwszych średnich z czynnikiem i+1, $L_i = \mu_1 + \mu_2 + ...\mu_i - i\mu_{i+1}$ (R: contr.hemlert(k))
- contr.poly: wielomiany ortogonalne, etc.

Przykład użycia kontrastów


```
contr = contr.sum(3)

model = lm(atrDec ~atr, data = dane, contrasts =
list(atr=contr))

summary(model)
```

Analiza wieloczynnikowa

W przeciwieństwie od analizy jednoczynnikowej, gdzie bada się średnie wartości atrybutu decyzyjnego y w zależności od jednego atrybutu nominalnego (kategorycznego), w analizie wieloczynnikowej bada się średnie wartości atrybutu decyzyjnego w zależności od kombinacji dwóch lub więcej atrybutów nominalnych.

Rozważa się:

- model addytywny (tzn. model bez interakcji)
- model z interakcjami

Analiza wieloczynnikowa a jednoczynnikowa

Przykład: cena mieszkania w zależności od:

- dzielnicy
- liczby pokoi
- typu budynku

W przypadku wielu atrybutów możliwe jest dokonanie wielokrotnej analizy jednoczynnikowej dla każdego atrybutu z osobna. Wtedy jednak mogą umknąć analizie wzajemne zależności między atrybutami. Aby tego uniknąć można dokonać analizy wieloczynnikowej.

Ograniczenia

Negatywne aspekty związane ze wzrostem liczby atrybutów:

- komplikacja modelu i analizy
- wykładniczy wzrost liczby możliwych kombinacji wartości
- spadek dokładności oceny efektów w modelu

Należy dopilnować, aby na każdą kombinację wartości atrybutów przypadało dostatecznie dużo obserwacji.

Model dwuczynnikowy

$$E(y|X = (a, b)) = \mu + \mu_a + \mu_b$$

μ jest wartością bazową, μ_a, μ_b to efekty wartości a i b dla dwóch analizowanych atrybutów, odpowiednio.

Zakłada się, że:

- $y|X \sim \mathcal{F}(\Theta)$, gdzie $\mathcal{F}$ to rodzina rozkładów normalnych z wariancją σ^2
- $\sum_a \mu_a = \sum_b \mu_b = 0$

Przykład w R

Symbolem używanym w R w definicji formuły w przypadku addytywnym jest '+':

```
anova(lm(cena~dzielnica+typ.budynku, data = flats))
```

Przykład analizy graficznej w R


```
plot.design(data.frame(flats$dzielnica, flats$typ.budynku,
flats$cena))
```

Analiza dwuczynnikowa z interakcją

Wartości atrybutów nie zawsze wpływają niezależnie i addytywnie na wartość atrybutu decyzyjnego.

Przed ewentualną analizą z interakcjami między atrybutami, można dokonać wstępnej analizy graficznej zależności.

Przykład graficzny w R

interaction.plot(dzielnica, typ.budynku, cena)

Jeśli wykresy są równoległe, świadczy to za addytywnością zależności. Brak równoległości świadczy za obecnością interakcji między atrybutami.

Model z interakcjami

$$E(y|X = (a, b)) = \mu + \mu_a + \mu_b + \mu_{ab}$$

Zakłada się, że:

- $\sum_a \mu_a = 0$
- $\sum_b \mu_b = 0$
- $\sum_a \mu_{ab} = 0$
- $\sum_b \mu_{ab} = 0$

Przykład w R

Warianty:

- tylko interakcje (R: symbol ':' w formule)
- interakcje oraz efekty addytywne (R: symbol '*' w formule)

```
anova(lm(cena dzielnica*typ.budynku, data = flats))
anova(lm(cena dzielnica + typ.budynku:dzielnica, data = flats))
```

Interakcje wyższych rzędów

Możliwe jest też uwzględnianie interakcji wyższych rzędów (tzn. wielomiany wyższych stopni wielu zmiennych), powinno się jednak używać takiego podejścia oszczędnie:

- problemy z interpretacją
- komplikacja modelu

Testy post-hoc

W przypadku analizy wieloczynnikowej można wykonywać testy post-hoc, jednak liczba kombinacji wartości artybutów rośnie bardzo szybko, co utrudnia interpretowalność.

Przykład:
TukeyHSD(aov(cena dzielnica+dzielnica:typ.budynku))

Wielowymiarowa analiza wariancji

W przypadku, gdy atrybut decyzyjny jest wielowymiarowy (np. wysokość przychodów w 3 różnych sektorach działalności, etc.) można dokonać analizy wielowymiarowej, która jest dość zaawansowanym zagadnieniem.

W R, jednym z narzędzi jest funcja `manova`. Można użyć wielu predefiniowanych testów, np. Pillai, Wilks, Hotelling-Lawley, etc.

Przykład: summary(manova(cbind(cena, powierzchnia) dzielnica+typ.budynku), test="Wilks")

Przykładowe pytania/zadania/problemy

- co to jest analiza wariancji i do czego służy
- jakie są rodzaje analizy wariancji
- typowe założenia w analizie wariancji
- jaki jest model w analizie jednoczynnikowej, dwuczynnikowej, z interakcjami
- podstawowe narzędzia w R do dokonania analizy wariancji

Dziękuję za uwagę.