Statystyczna analiza danych (molekularnych) – analiza wariancji ANOVA

Anna Gambin

19 maja 2013

Spis treści

1 Przykład: Model liniowy dla ekspresji genów

Na poprzednich wykładach omawialiśmy T-test, który potrafi wskazać geny o zróżnicowanej ekspresji w dwóch grupach pacjentów. Nasuwa się oczywiste uogólnienie, co czynić jeśli interesują nas geny różnicujące trzy populacje. Pomoże nam w tym zadaniu odpowiedni model liniowy i technika zwana analizą wariancji (ANOVA). Jak zwykle poprawność naszego rozwiązania jest warunkowana założeniem, że poziomy ekspresji badanych genów mają rozkład normalny o identycznej wariancji we wszystkich grupach.

Niech zmienna Y_i oznacza oznacza poziom ekspresji. Rozważamy k grup pacjentów. Zakładamy następujący model liniowy:

$$Y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \epsilon_i$$

gdzie $x_{ij} = 1$ jeśli pacjent i-ty należy do grupy j-tej i $x_{ij} = 0$ wpp. W powyższym modelu rozpoznajemy omawiany już model bez predyktorów, dla którego parametry $\beta_1, \beta_2, \ldots \beta_k$ odpowiadają wartościom średnim w grupach.

Załóżmy, że chcemy przetestować hipotezę zerową mówiąca, że średnia ekspresja danego genu w trzech lub więcej rozważanych grupach jest równa, czyli $H_0 : \mu_1 = \mu_2 = \mu_3$. Niech pomiary ekspresji badanego genu w pierwszej grupie będą oznaczane jako $y_{11}, y_{21}, \ldots y_{n1}$, w drugiej grupie odpowiednio $y_{12}, y_{22}, \ldots, y_{n2}$, i analogicznie w trzeciej grupie $y_{13}, y_{23}, \ldots, y_{n3}$. Policzmy średnie ekspresje genu w grupach:

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ji}, \qquad \text{dla} \quad i = 1, 2, 3$$

Niech $\bar{y}$ oznacza średnią ekspresje we wszystkich grupach, czyli

$$\bar{y} = \frac{1}{3}(\bar{y}_1 + \bar{y}_2 + \bar{y}_3)$$

Policzymy teraz dwie sumy kwadratów odchyleń od średniej: wewnątrz grup (SSW) i pomiędzy grupami (SSB):

$$\mathrm{SSW} = \sum_{i=1}^{3}\sum_{j=1}^{n}(y_{ji} - \bar{y}_i)^2,$$

$$\mathrm{SSB} = \sum_{i=1}^{3}\sum_{j=1}^{n}(\bar{y}_i - \bar{y})^2 = n\sum_{i=1}^{3}(\bar{y}_i - \bar{y})^2.$$

Zdefiniujemy teraz f-statystykę jako:

$$f = \frac{\mathrm{SSB}\,/(3-1)}{\mathrm{SSW}\,/(3n-3)}$$

jeśli rozważamy k grup statystyka jest równa:

$$f = \frac{\mathrm{SSB}\,/(k-1)}{\mathrm{SSW}\,/(kn-k)}$$

Przy założeniu, że dane pochodzą z rozkładu normalnego f-statystyka ma rozkład $F_{k-1,kn-k}$, czyli rozkład F o $(k-1, kn-k)$ stopniach swobody. Możemy teraz odrzucić hipotezę zerową jeżeli $P(F_{k-1,kn-k} > f) < \alpha$. Intuicyjnie przy założeniu hipotezy zerowej (jeśli średnie w grupach są równe) to wartość statystyki SSB powinna być mała, podobnie jak wartość f-statystyki, co spowoduje przyjęcie H_0.

2 Jednoczynnikowa analiza wariancji

Analiza wariancji (w skrócie ANOVA) jest bardzo ważną techniką, której zastosowanie widzieliśmy w ostatnim przykładzie. Używana jest w wielu zagadnieniach, w bioinformatyce służy najczęściej porównywaniu średnich w wielu grupach, ale nie tylko. Analiza wariancji została stworzona w latach dwudziestych ubiegłego wieku przez Ronalda Fishera.

Załóżmy, że dysponujemy modelem liniowym dla zbioru obserwacji. Przyjęło się w kontekście analizy wariancji nazywać zmienne objaśniające, czyli predyktory **czynnikami**, natomiast parametry będziemy nazywać **efektami**. Naszym celem jest wyodrębnić w całkowitej wariancji odpowiedzi Y, składniki pochodzące od poszczególnych czynników, oraz wariancję, za którą odpowiedzialny jest błąd.

Oznacza to, że wariancja w danych może mieć zarówno przyczyny identyfikowalne (wtedy można próbować ja zmniejszyć, bo mamy na nią wpływ) oraz przyczyny pozostające poza naszą kontrolą.

Analiza wariancji dostarcza informacji potrzebnych do wnioskowania na temat średnich w grupach: jeśli średnie w grupach się znacząco różnią możemy odrzucić hipotezę zerową zakładających ich równość, o ile wariancja w każdej próbie jest odpowiednio mała w odniesieniu do całkowitej wariancji.

Sytuacja, w której wariancja w grupach jest duża w porównaniu z całkowitą wariancją nie pozwala nam na odrzucenie hipotezy zerowej. Podstawowe założenia pozwalające stosować F-test w powyższym przykładzie i ogólnie w analizie wariancji to:

- wszystkie obserwacje są niezależne,
- pochodzą z populacji o rozkładach normalnych,
- rozważane efekty są addytywne.

Ze względu na wymóg normalności rozkładów w badanych grupach możemy zaliczyć technikę ANOVA do testów parametrycznych.

Sformułujemy teraz w pełnej ogólności metodę analizy wariancji dla jednego czynnika (predyktora).

Rozważmy N obserwacji Y_{ij} gdzie $i = 1, 2 \ldots k$ oraz $j = 1, 2 \ldots n_i$. Zmienna (właściwie próba) losowa Y jest pogrupowana w k klas o licznościach $n_1, n_2, \ldots n_k$, $N = \sum_{i=1}^{k} n_i$. Możemy w kontekscie medycyny molekularnej myśleć o obserwacjach ekspresji genu w różnych tkankach.

Oznaczmy:

$$\bar{Y_{i.}} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

$$T_{i.} = \sum_{j=1}^{n_i} Y_{ij}$$

$$G = \sum_{i=1}^{k} T_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

Często prezentujemy dane do analizy w postaci tabeli:

grupa	obserwacje	średnie	sumy
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\bar{Y_1}$	T_1
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\bar{Y_2}$	T_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$	$\bar{Y_k}$	T_k
			G

Jeżeli badane klasy są równoliczne, czyli zachodzi $n_1 = n_2 = \ldots = n_k$, to mamy doczynienia z przypadkiem **zrównoważnonym**.

Ponownie rozważmy model liniowy ($i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$):

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

gdzie μ nazywane ogólnym efektem średnim wynosi:

$$\mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{N}$$

natomiast μ_i nazywamy efektem i-tej klasy (grupy). Zakładamy dodatkowo, że błąd ϵ_{ij} ma rozkład normalny o średniej zero i ustalonej (niezależnej od klasy) wariancji σ_ϵ^2.

Przyjmijmy, że chcemy przetestować hipotezę mówiącą, że średni efekt dla wszystkich tkanek jest taki sam, czyli mamy dwie równoważne hipotezy zerowe:

$$H_0 : \ \mu_1 = \mu_2 = \ldots = \mu_k = \mu \tag{1}$$

$$H_0 : \ \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0 \tag{2}$$

Aby znaleźć μ i α_i stosujemy metodę najmniejszych kwadratów:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \epsilon_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

Liczymy odpowiednie pochodne cząstkowe i przyrównujemy do zera:

$$\frac{\partial E}{\partial \mu} = 0; \quad \frac{\partial E}{\partial \alpha_i} = 0 \quad \forall 1 \leq i \leq k$$

i otrzymujemy

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} 1} = \bar{Y}_{..} \text{ oraz } \hat{\alpha}_i = \bar{Y}_{i.} - \hat{\mu}$$

Policzmy sumę kwadratów odchyleń od wartości średnich, czyli wartość TSS (**T**otal **S**um of **S**quares):

$$\text{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.} + \bar{Y}_{i.} - \bar{Y}_{..})^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + 2[\sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{Y}_{..}) \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})]$$

Zauważmy, że

$$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.}) = 0$$

czyli podsumowując:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

Składnik $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$ nazwiemy SSE jako sumę kwadratów błędów (**S**um of **S**quares of **E**rrors), a składnik $\sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$ określimy jako SST, czyli **S**um of **S**quares due to **T**reatment. Używając wprowadzonych oznaczeń nasza zależność jest następująca:

$$\text{TSS} = \text{SSE} + \text{SST}$$

Zauważmy, że licząc statystykę TSS wykorzystujemy N zmiennych przy dodatkowym ograniczeniu:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..}) = 0$$

Wnioskujemy stąd, że ma ona $(N-1)$ stopni swobody. Podobnie dla SST mamy $(k-1)$ stopni swobody, gdyż dysponujemy k obserwacjami i dodatkowym warunkiem:

$$\sum_{i=1}^{k} n_i(\bar{Y_{i.}} - \bar{Y_{..}}) = 0$$

Statystyka SST ma $(N-k)$ stopni swobody, gdyż jest liczona z użyciem N obserwacji podlegających k ograniczeniom:

$$\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y_{i.}}) = 0 \quad i = 1, 2, \ldots k$$

W celu testowania hipotez zerowych (1) i (2) użyjemy F-testu, który bada czy wariancje w dwóch grupach są równe. Jako średnią sumę kwadratów przyjmiemy odpowiednią sumę kwadratów podzieloną przez liczbę stopni swobody, czyli

$$\text{MST} = \frac{SST}{k-1}$$

$$\text{MSE} = \frac{SSE}{N-k}$$

gdzie MST jest skrótem od *Mean Sum of Squares due to Treatment*, a MSE oznacza *Mean Sum of Squares*. Dwie wprowadzone wielkości szacują wariancję w grupach i ogólną wariancję w danych, a ich iloraz ma rozkład F o $k-1$ i $N-k$ stopniach swobody:

$$F = \frac{MST}{MSE} \sim F_{k-1,N-k}$$

2.1 Testy *post-hoc*

Zauważmy, że test ANOVA pozwala jedynie odrzucić hipotezę zerową o równości średnich w grupach. Nie wskazuje jednak, które średnie znacząco różnią się między sobą. Dla znalezienia takich grup stosuje się testy typu *post-hoc*.

Do takich testów należą m.in.:

- test HSD Tukeya (HSD - **H**onestly **S**ignificant **D**ifference);
- test Studenta-Newmana-Keulsa;
- test LSD Fishera (LSD- **L**east **S**ignificant **D**ifference).

3 Wieloczynnikowa analiza wariancji

W poprzednim rozdziale badaliśmy relacje pomiędzy grupami obiektów określonymi przy pomocy jednej zmiennej jakościowej (która indukowała podział na kategorie). Teraz założymy, że badane zagadnienie opisują dwie zmienne jakościowe (przedstawioną metodę można uogólnić na więcej zmiennych).

Mamy N obserwacji oraz dwie zmienne jakościowa A (występuje na k poziomach) oraz zmienna B (występuje na h poziomach). Zakładamy, że nasze obserwacje pochodzą z rozkładu normalnego o średnich specyficznych dla danej grupy (o liczności n_{ij}, wyznaczonej przez zmienne A i B):

$$Y_{ijm} \sim N(\mu_{ij}, \sigma^2) \quad 1 \leq i \leq k, q \leq j \leq h, 1 \leq m \leq n_{ij}$$

Rozważmy model liniowy:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \quad 1 \leq i \leq k, 1 \leq j \leq h$$

gdzie μ jest nazywane ogólnym efektem średnim:

$$\mu = \frac{\sum_{i=1}^{k} \sum_{j=1}^{h} \mu_{ij}}{N}$$

α i β to addytywne efekty zmiennych, natomiast γ_{ij} opisuje efekt interakcji zmiennych w bloku (i, j), a błąd losowy $\epsilon \sim N(0, \sigma^2)$.

Oznaczmy:

$$\mu_{i.} = \frac{\sum_{i=1}^{h} \mu_{ij}}{h}$$

$$\mu_{.j} = \frac{\sum_{i=1}^{k} \mu_{ij}}{k}$$

Mamy wtedy:

$$\alpha_i = \mu_{i.} - \mu$$

$$\beta_j = \mu_{.j} - \mu$$

$$\gamma_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$$

Zwróćmy uwagę, że występowanie interakcji nie oznacza, że model przestaje być liniowy. Ażeby opis był jednoznaczny potrzebne są ograniczenia na parametry:

$$\sum_{i=1}^{k} \alpha_i = 0 \text{ oraz } \sum_{j=1}^{h} \beta_j = 0$$

Zajmijmy się na początek modelem bez interakcji (wpływy zmiennych A i B są niezależne). Naszym zadaniem jest przetestowanie następujących hipotez zerowych:

$$H_0 : \ \mu_{1.} = \mu_{2.} = \ldots = \mu_{k.} = \mu \tag{3}$$
$$: \quad \mu_{.1} = \mu_{.2} = \ldots = \mu_{.k} = \mu \tag{4}$$

lub równoważnie

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0 \tag{5}$$
$$: \ \beta_1 = \beta_2 = \ldots = \beta_h = 0 \tag{6}$$

Zastosujemy po raz kolejny metodę najmniejszych kwadratów:

$$E = \sum_{i=1}^{k} \sum_{i=1}^{h} \epsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} (Y_{ij} - \mu - \alpha_i - \beta_j)^2$$

Różniczkujemy po μ, α_i i β_j i przyrównujemy do zera otrzymując estymatory:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{h} Y_{ij}}{N} = \bar{Y}_{..}$$
$$\hat{\alpha}_i = \frac{\sum_{j=1}^{h} Y_{ij}}{h} - \hat{\mu} = \bar{Y}_{i.} - \bar{Y}_{..}$$
$$\hat{\beta}_j = \frac{\sum_{i=1}^{k} Y_{ij}}{k} - \hat{\mu} = \bar{Y}_{.j} - \bar{Y}_{..}$$

Zajmiemy się teraz dekompozycją zmienności w danych. Mamy

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{h} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} (Y_{ij} - \bar{Y}_{i.} + \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$= h \sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + k \sum_{j=1}^{h} (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{h} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} - \bar{Y}_{..})^2$$

czyli $TSS = SST + SSB + SSE$

Podobnie jak poprzednio:

$$SST = h \sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SSB = k\sum_{j=1}^{h}(\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{h}(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} - \bar{Y}_{..})^2$$

Testowanie hipotez zerowych (3) i (4) (lub równoważnie (5) i (6)) opiera się na następujących statystykach o rozkładzie F.

$$\frac{MST}{MSE} \sim F_{k-1,(k-1)(h-1)} \text{ oraz } \frac{MSB}{MSE} \sim F_{h-1,(k-1)(h-1)}$$

gdzie średnie sumy kwadratów odchyleń otrzymujemy dzieląc przez odpowiednią liczbę stopni swobody.

$$MST = \frac{SST}{k-1}, \quad MSB = \frac{SSB}{h-1}, \quad MSE = \frac{SSE}{(k-1)(h-1)}$$

Literatura

[1] Julian J. Faraway Practical Regression and Anova using R, 2002.

[2] K. Seefeld, E. Linder, Statistics using R with biological examples, 2007.

[3] W.P. Krijnen Applied statistics for bioinformatics using R, 2009.

[4] S.K. Mathur, Statistical Bioinformatics with R, Elsevier Academic Press, 2010.

[5] W. J. Ewens, G. R. Grant, Statistical Methods in Bioinformatics Springer-Verlag, 2001.