Matematyczne metody w naukach biomedycznych: regresja i analiza wariancji.

Anna Gambin

23 listopada 2013

Spis treści

1 Analiza regresji

Analiza regresji jest używana modelowania związków zachodzących pomiędzy zmienną losową Y (zwaną często odpowiedzią, zmienną zależną lub *zmienną objaśnianą*), a jedną lub więcej *zmiennymi objaśniającymi* (zwanymi też *predyktorami*): $X_1, X_2, \ldots X_p$. Dla $p = 1$ mamy do czynienia z prostą regresją, natomiast dla $p > 1$ mówimy o *regresji wielorakiej*.

Najczęściej przyjmujemy, że odpowiedź Y jest zmienną ciągłą, natomiast zmienne $X_1, X_2, \ldots X_p$ mogą być zarówno ciągłe jak i dyskretne. Podstawowe cele jakie stawiamy sobie w zadaniu regresji to:

- przewidywanie przyszłych obserwacji;
- opisanie wpływu predyktorów na odpowiedź Y;
- uzyskanie ogólnego opisu struktury posiadanych danych.

Istnieje wiele uogólnień zadania regresji, których nie będziemy rozważać, zainteresowanych odsyłam do szerokiej literatury (również odpowiednich pakietów R-owych: `http://cran.r-project.org/`). Wspomniane uogólnienia to: dopuszczenie wielu odpowiedzi $Y_1, Y_2, \ldots Y_k$, przypadek w którym odpowiedź Y jet binarna (*regresja logistyczna*), oraz regresja Poissona, gdzie zakłada się, że odpowiedź Y przyjmuje wartości całkowite nieujemne.

1.1 Historia

Po raz pierwszy zadanie regresji zostało sformułowane w XVIII wieku w kontekście zastosowań wiedzy astronomicznej do nawigacji. Metoda najmniejszych kwadratów (którą przedstawimy poniżej) została zaproponowana przez francuskiego matematyka Adrien-Marie Legendre w roku 1805. Książe matematyków Carl Friedrich Gauss utrzymywał, że opracował tę metodę jeszcze wcześniej i w roku 1809 udowodnił, że najmniejsze kwadraty prowadzą do rozwiązania optymalnego przy założeniu, że błędy mają rozkład normalny.

2 Modele liniowe

Zacznijmy od rozważenia (zupełnie nie bioinformatycznego) przykładu: niech zmienna objaśniana Y opisuje zużycie paliwa. Jako predyktory (zmienne objaśniające) posłużą nam dane o X_1, czyli wadze pojazdu, X_2 – mocy silnika, oraz X_3 – liczbie cylindrów. Załóżmy, że dysponujemy n obserwacjami (przypadkami):

$$\begin{array}{cccc} y_1 & x_{11} & x_{12} & x_{13} \\ y_2 & x_{21} & x_{22} & x_{23} \\ \ldots & & \ldots & \\ y_n & x_{n1} & x_{n2} & x_{n3} \end{array}$$

Najbardziej ogólna postać modelu opisującego nasze dane mogłaby być następująca:

$$y = f(X_1, X_2, X_3) + \epsilon$$

gdzie f jest pewną nieznaną funkcją, ϵ addytywnym błędem. Zazwyczaj nie posiadamy wystarczającej liczby danych, żeby podjąć próbę estymacji "dowolnie dzikiej" funkcji f i dlatego przyjmujemy upraszczające założenia co do klasy rozważanych funkcji. Bardzo prostym, a jednocześnie wystarczająco elastycznym założeniem jest, że ma ona postać liniową:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Stałe β_i nazywamy *parametrami*, i całe zadanie zbudowania modelu liniowego sprowadza się do ich estymacji z danych. Zauważmy, że założenie liniowości dotyczy parametrów modelu, natomiast zmienne predyktory mogą występować w dowolnej postaci. Np model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \log X_2 + \beta_3 \log X_3 + \epsilon$$

pozostaje modelem liniowym, natomiast model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\beta_3} + \epsilon$$

nie jest liniowy ze względu na parametry.

Wróćmy do naszego samochodowego przykładu (oczywiście, jak bardzo nam zależy możemy zamiast o zużyciu paliwa myśleć o niezbędnej dawce leku, a parametry pojazdu zamienić na dane o stanie pacjenta). Dla $i = 1, 2, \ldots n$ możemy zapisać:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

W postaci macierzowej równanie regresji wygląda bardziej przejrzyście:

$$y = X\beta + \epsilon$$

gdzie $y = (y_1, \ldots y_n)^T$, $\epsilon = (\epsilon_1, \ldots \epsilon_n)^T$, $\beta = (\beta_0, \ldots \beta_3)^T$ oraz

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \ldots & & \ldots & \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}$$

Przykład 1 *Zapiszmy w notacji macierzowej bardzo prosty model, w którym nie występują żadne zmienne objaśniające: $y = \mu + \epsilon$, czyli obserwacje opisujemy za pomocą wartości średniej oraz błędu (zmiennej losowej o zerowej wartości oczekiwanej).*

$$\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \ldots \\ \epsilon_n \end{pmatrix}$$

Przykład 2 *Załóżmy, że badamy odpowiedź na określoną terapię w dwóch rozłącznych grupach pacjentów. W pierwszej grupie odpowiedzi oznaczamy $y_1, \dots y_m$ a średnia odpowiedź wynosi μ_y, natomiast w drugiej grupie obserwujemy odpowiedzi $z_1, \dots z_n$ ze średnią μ_z. Macierz zmiennych objaśniających jest w tym przypadku zero-jedynkowa i wskazuje przynależność danego pacjenta do grupy.*

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \\ z_1 \\ z_2 \\ \dots \\ z_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \dots \\ \dots \\ \epsilon_{n+m} \end{pmatrix}$$

3 Estymacja najmniejszych kwadratów

Zastanówmy się teraz jak estymować parametry β_i. Naszym zadaniem jest możliwie najdokładniejsza reprezentacja danych wymiaru n za pomocą modelu wymiaru p. Oczywiście pozostaną jeszcze losowe fluktuacje, czyli błąd. Geometrycznie możemy wyobrazić sobie optymalny model jako rzut ortogonalny wektora y (leżącego w przestrzeni n-wymiarowej na przestrzeń p wymiarową rozpinaną przez predyktory.

Pozostawiając na chwile intuicję geometryczną (która *nota bene* jest zbieżna z poniższą metodą) załóżmy, że estymując β chcemy minimalizować sumę kwadratów błędów, czyli $\epsilon^T\epsilon$. Szukamy wektora $\hat{\beta}$ minimalizującego:

$$R = \sum_i \epsilon_i^2 = \epsilon^T\epsilon = (y - X\beta)^T(y - X\beta)$$

Przekształcay powyższe, różniczkujemy ze względu na β i przyrównujemy do zera:

$$R = y^Ty - 2\beta X^Ty + \beta^T X^T X\beta$$

$$\frac{\partial R}{\partial \beta} = 0 \implies X^TX\hat{\beta} = X^Ty$$

Przy założeniu, że macierz X^TX jest odwracalna mamy:

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

Zauważmy, że $x\hat{\beta} = X(X^TX)^{-1}X^Ty$ jest po prostu rzutem ortogonalnym y na przestrzeń rozpinaną przez X. Macierz $H = X(X^TX)^{-1}X^T$ jest macierzą rzutu prostopadłego, zgrabnym obiektem teoretycznych rozważań, jednak zupełnie nieprzydatnym w praktyce ze względu na jej wymiar n x n.

Podsumujmy:

- wartości przewidywane przez model: $\hat{y} = Hy = X\hat{\beta}$.
- residua modelu: $\hat{\epsilon} = y - X\hat{\beta} = y - \hat{y} = (I - H)y$.
- suma kwadratów residuów: $\hat{\epsilon}^T\hat{\epsilon} = y^T(I - H)y$.

Przykład 3 *Policzmy teraz $\hat{\beta}$ dla prostego modelu bez zmiennych objaśniających: Mamy $y = \mu + \epsilon$, czyli $X = \mathbf{1}$ i $\beta = \mu$, więc $X^TX = \mathbf{1}^T\mathbf{1} = n$. Dalej*

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \frac{1}{n}\mathbf{1}^Ty = \bar{y}$$

4 Twierdzenie Gaussa Markowa: optymalność $\hat{\beta}$

Podamy teraz argumenty za tym, że $\hat{\beta}$ uzyskane metodą najmniejszych kwadratów jest sensownym oszacowaniem dla parametrów modelu liniowego. Po pierwsze jak pokazaliśmy, odpowiada ono rzutowi prostopadłemu na przestrzeń modelu. Po drugie jeśli błędy są niezależne i pochodzą z rozkładu normalnego estymator uzyskany metodą najmniejszych kwadratów jest tożsamy z estymatorem największej wiarygodności, czyli jest takim oszacowaniem parametrów, które maksymalizuje prawdopodobieństwo zaobserwowania danych przy założeniu modelu. Jako trzeci argument przytoczymy twierdzenie Gaussa Markowa, które mówi, że jest to najlepszy (**B**EST), liniowy (**L**INEAR), nieobciążony (**U**NBIASED) **E**stymator (w angielskim sktócie: BLUE).

Do sformułowania twierdzenia potrzebujemy pojęcia *funkcji estymowalnej*. Powiemy, że liniowa kombinacja parametrów $\Psi = c^T\beta$ jest estymowalna wtedy i tylko wtedy jeśli istnieje liniowa kombinacja a^Ty, taka że:

$$E(a^Ty) = c^T\beta$$

Twierdzenie: Załóżmy teraz, że wartość oczekiwana błędu wynosi zero: $E(\epsilon) = 0$ oraz wariancja var$(\epsilon) = \sigma^2 I$. Dodatkowo załóżmy, że strukturalna część modelu $E(y) = X\beta$ jest skonstruowana poprawnie. Niech $\Psi = c^T\beta$ będzie estymowalną funkcją, wtedy spośród wszystkich nieobciążonych liniowych estymatorów Ψ, estymator najmniejszych kwadratów $\hat{\beta}$ ma najmniejszą wariancję i jest wyznaczony jednoznacznie.

dowód: Załóżmy, że $a^T y$ jest pewnym nieobciążonym estymatorem $c^T \beta$, czyli

$$E(a^T y) = c^T \beta = a^T X \beta \qquad \forall \beta$$

Wnioskujemy stąd, że $a^T X = c^T$, czyli wektor c leży w przestrzeni rozpiętej przez kolumny X^T, a więc również przez $X^T X$. Oznacza to, że istnieje wektor współczynników λ, pozwalający wyrazić c jako kombinację liniową wektorów z bazy:

$$c = X^T X \lambda$$

$$c^T \hat{\beta} = \lambda^T X^T X \hat{\beta} = \lambda^T X^T y$$

Pokażemy teraz, że nasz estymator ma najmniejszą wariancję. Weźmy dowolny estymator $a^T y$ i policzmy jego wariancję:

$$\text{var}(a^T y) = \text{var}(a^T y - c^T \hat{\beta} + c^T \hat{\beta}) = \text{var}(a^T y - \lambda^T X^T y + c^T \hat{\beta}) =$$

$$= \text{var}(a^T y - \lambda^T X^T y) + \text{var}(c^T \hat{\beta}) + 2\text{cov}(a^T y - \lambda^T X^T y, \lambda^T X^T y)$$

Można pokazać, że $\text{cov}(a^T y - \lambda^T X^T y, \lambda^T X^T y) = 0$[1], czyli

$$\text{var}(a^T y) = \text{var}(a^T y - \lambda^T X^T y) + \text{var}(c^T \hat{\beta})$$

Ponieważ wariancje są nieujemne, widzimy, że $\text{var}(a^T y) > \text{var}(c^T \hat{\beta})$. Pozostaje jeszcze wykazać jednoznaczność estymatora $\hat{\beta}$. Mamy $\text{var}(a^T y) = \text{var}(c^T \hat{\beta})$ jedynie dla $\text{var}(a^T y - \lambda^T X^T y) = 0$, czyli $a^T y - \lambda^T X^T y = 0$, co oznacza, że $a^T y = \lambda^T X^T y = c^T \hat{\beta}$.

Powyższe twierdzenie uzasadnia, że estymator najmniejszych kwadratów jest dobrym wyborem, jednak jeśli błędy są skorelowane, albo mają różnicowaną wariancję mogą istnieć lepsze estymatory; podobnie jeśli błędy nie pochodzą z rozkładu normalnego.

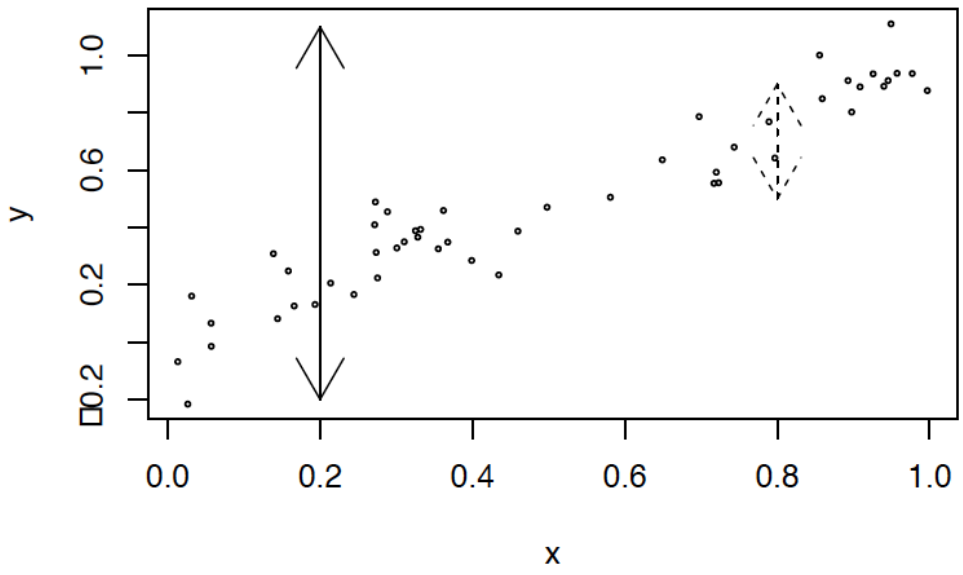
Policzymy jeszcze wartość oczekiwaną wariancję dla $\hat{\beta}$.

$$E(\hat{\beta}) = (X^T X)^{-1} X^T X \beta = \beta$$

czyli jak obiecywaliśmy, estymator jest nieobciążony.

$$\text{var}(\hat{\beta}) = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2$$

[1]Przypomnijmy, że kowariancja dla dwóch zmiennych losowych X i Y jest zdefiniowana następująco: $\text{cov}(X, Y) = \text{E}(X \cdot Y) - \text{E}X \cdot \text{E}Y$



Rysunek 1: Strzałka kropkowana pokazuje rozrzut odpowiedzi y pod warunkiem, że znamy x, natomiast strzałka ciągła odpowiada zmienności odpowiedzi y, jeśli nie znamy zmiennej objaśniającej x.

4.1 Jakość dopasowania

Jako miarę jakości dopasowania modelu do danych rozważmy współczynnik zwany *procentem wyjaśnionej wariancji*:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$$

Współczynnik R^2 przyjmuje wartości od 0 do 1, przy czym wartości bliższe 1 odpowiadają lepszemu dopasowaniu modelu do danych. Intuicja stojąca za jego definicją jest następująca: załóżmy, że chcemy przewidzieć odpowiedź y. Jeśli nie znamy x to najlepszą predykcją będzie $\bar{y}$, ale jej rozrzut będzie duży. Jeżeli natomiast znamy x to za najlepszą predykcję uznamy dopasowanie przy pomocy regresji, czyli $\hat{y}$. Jeżeli istnieje jakaś systematyczna zależność między x i y to zmienność (rozrzut) takiej predykcji powinna być mniejsza (por. Rys. 1). Współczynnik R^2 wynosi 1 minus stosunek sum kwadratów dla tych obydwu predykcji. Jeżeli dopasowanie regresyjne jest idealne, to ten stosunek wynosi 0, a procent wyjaśnionej wariancji 1.

5 Przykład: Model liniowy dla ekspresji genów

Poprzednio 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:

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

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.

6 Jednoczynnikowa analiza wariancji (ANOVA)

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:

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

6.1 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).

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.